PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : C12Q 1/68	A2	(11) International Publication Number: **WO 00/18966**
		(43) International Publication Date: 6 April 2000 (06.04.00)

(21) International Application Number: PCT/US99/22975

(22) International Filing Date: 29 September 1999 (29.09.99)

(30) Priority Data:
60/102,381 29 September 1998 (29.09.98) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 60/102,381 (CON)
Filed on 29 September 1998 (29.09.98)

(71) Applicant *(for all designated States except US)*: ARCH DEVELOPMENT CORPORATION [US/US]; 5640 South Ellis Avenue, Chicago, IL 60637 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: WANG, San, Ming [-/US]; - (US). FEARS, Scott [-/US]; - (US). ROWLEY, Janet, D. [-/US]; - (US).

(74) Agent: HIGHLANDER, Steven, L.; Arnold, White & Durkee, P.O. Box 4433, Houston, TX 77210 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: A NEW STRATEGY FOR GENOME-WIDE GENE ANALYSIS: INTEGRATED PROCEDURES FOR GENE IDENTIFICATION

(57) Abstract

The present invention is drawn to methods for genome-wide index in a particular cell type, and for genome-wide identification of differentially expressed genes. Also provided is an improved set of poly-dT primers anchored at their 3' ends.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

DESCRIPTION

A NEW STRATEGY FOR GENOME-WIDE GENE ANALYSIS: INTEGRATED PROCEDURES FOR GENE IDENTIFICATION

BACKGROUND OF THE INVENTION

The present application is a continuation–in–part of co–pending U.S. Patent Application Serial No. 60/102,381, filed September 29, 1998. The entire text of the above–referenced disclosure is specifically incorporated by reference herein without disclaimer. The government may own rights in the present invention pursuant to grant number CA42557 from The National Institutes of Health.

A. Field of the Invention

The present invention relates to the fields of molecular biology and nucleic acid biochemistry. More particularly, the invention provides new methods for genome-wide gene identification.

B. Related Art

The rapid progress of the human genome project allows new strategies for the functional genomic analysis of normal and abnormal cells. The total number of expressed human genes has been estimated to be about 100,000, with about 11,000 genes being expressed in any particular cell type (Alberts *et al.*, 1994). These genes can be grouped by their level of expression into abundant, intermediate abundant and rare abundant classes. These classes contain about 4-10 genes, 500 genes, and 11,000 genes respectively, comprising 10%, 40%, and 50% of the total transcripts (Alberts *et al.*, 1994). The majority of expressed genes, therefore, belong to the rare abundant class. Most of the processes for gene identification also need to focus on this category.

Gene expression is tightly regulated in both temporal and tissue specific fashions. Abnormal gene expression in pathological situations can alter the normal cellular behavior leading to various abnormalities such as neoplasia. Analysis of gene

expression in various normal conditions can provide information regarding basic cell physiology. In pathological conditions, the abnormally expressed genes can serve as markers for early diagnosis, as targets for drug design, as indicators for treatment responsiveness, and for prognosis.

Over one million expressed sequence tags (EST) from the human genome are listed in the current NCBI dbEST database. Ultimately, most of the expressed genes from human genome will be indexed in the EST database. Maximal use of EST information will greatly accelerate the gene identification process, *e.g.*, using an EST sequence to search the UniGene database to obtain the cluster information for that sequence and to obtain the original plasmids used for EST project for further analysis (Boguski, 1995; Gerhold and Caskey, 1996).

Due to the large size of the human genome, there is a critical need to have powerful techniques for genome-wide gene identification. Most of the methods currently used for gene identification at the whole genome level can be performed in only a few laboratories because they are either complicated or very costly. These include DNA microarray techniques (Lockhart *et al.*, 1996; DeRisi *et al.*, 1996), serial analysis of gene expression technique (SAGE) (Velculescu *et al.*, 1995), or large-scale sequencing in the cancer genome anatomy project (CGAP) (Strausberg *et al.*, 1997).

Differential Display technique is another powerful technique for analyzing expressed genes (Liang and Pardee, 1992). Differential Display compares the expressed sequence profiles of different cell types by a combination of reverse transcription, using sets of 3′-anchored oligo-dT and 5′ arbitrary primers, along with PCR™ amplification. Separation of the amplified products reveals unique banding patterns that identify differences in the expressed products between two mRNA sources. However, this technique gives a high frequency of false positives. Suppression Subtraction Hybridization (SSH) is another technique that recently has been developed (Diatchenko *et al.*, 1996). Though this technique requires much less mRNA than the classical subtraction technique, and can enrich some particular transcripts by a thousand-fold (Diatchenko *et al.*, 1996), the uncertainty of the sequence location within the gene due to lack of selection of specific regions of the

gene means that many sequences obtained from SSH cannot be matched to the EST sequences as they are not located in the 3′ or 5′ portion of the gene. These sequences are incorrectly considered to be "novel sequences" representing potential novel genes (Diatchenko *et al.*, 1996).

It is important to develop new, more efficient techniques to pursue genome - wide analysis. These techniques should be highly sensitive in order to identify the rare templates, should require a relatively small amount of initial mRNA, and should be largely based on routine molecular biology techniques, which would be suitable for use in a standard research laboratory, and more importantly, should maximally use the publicly accessible databases for gene identification.

SUMMARY OF THE INVENTION

The present invention provides a method for amplifying a first set of target polynucleotides containing poly-A sequences comprising (a) providing a set of five primers, wherein each of the primers is comprised a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC; (b) annealing the primers to the first set of target polynucleotides; (c) contacting the primer-annealed first set of target polynucleotides with a polymerase and dNTPs; and (d) subjecting the components of step (c) to conditions permitting polymerization, whereby a first set of polymerization products is generated.

In particularly preferred embodiments, the polymerase is a reverse transcriptase. In more specific embodiments, the reverse transcriptase is MMLV reverse transcriptase. In other particular embodiments, the polymerase is a DNA polymerase. More particularly, in preferred embodiments, the DNA polymerase is Taq. In preferred embodiments, the primers further comprise a sequence encoding a promoter 5′ to the poly-dT sequence. More particularly, the promoter may be an SP6 promoter, an M13 promoter, a T3 promoter or a T7 promoter. In certain embodiments, the invention may further comprise subjecting the first set of polymerization products to PCR. In other embodiments, the method may further

comprise separating the first set of polymerization products. More particularly, the separating comprises gel electrophoresis. In preferred embodiments, the gel electrophoresis comprises denaturing gel electrophoresis.

In specific aspects of the invention, the poly-dT sequence is about 10 to about 35 bases. In more particular embodiments, the poly-dT sequence is 11 bases. In specific embodiments, the primers contain a label. More particularly, the label may be a fluorometric label, colorimetric label, enzymatic label, magnetic label, biotin label or radioactive label. In other specific embodiments, the target polynucleotide may be an RNA or a DNA.

In certain defined embodiments of the present invention, the first set of polymerization products are compared with a second set of polymerization products generated from a second set of target polynucleotides. In specific embodiments, the comparison is differential display.

Also provided herein is a method for generating DNA library from poly-A RNAs comprising (a) providing a set of five primers, wherein each of the primers is comprised a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC; (b) annealing the primers to the RNAs; (c) contacting the primer annealed set of target polynucleotides with a polymerase dNTPs; (d) subjecting the components of step (c) to conditions permitting polymerization; and (e) cloning polymerization products of step (d) into a suitable vector; whereby a DNA library is generated.

Specifically, the primers further may comprise a sequence encoding a promoter 5′ to the poly-dT sequence. In particularly preferred embodiments, the promoter is a SP6 promoter, a T3 promoter or a T7 promoter. In other preferred embodiments, the vector is an expression vector. In certain aspects of the present invention, polymerase is a reverse transcriptase. In other aspects of the method further may comprise subjecting the polymerization products to PCR.

The present invention further describes a method for performing differential display comprising (a) providing a set of five primers, wherein each of the primers is comprised a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic

acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC; (b) annealing the primers to a first set of target polynucleotides containing poly-A sequences; (c) contacting the primer-annealed first set of target polynucleotides with a polymerase and dNTPs; (d) subjecting the components of step (c) to conditions permitting polymerization to create a first set of polymerization products; and (e) comparing the first set of polymerization products with a second set of polymerization products produced according to steps (a)-(d) using a second set of target polynucleotides containing poly-A sequences.

In specific embodiments, the method further comprises subjecting the first set of polymerization products to PCR.

Also described by the present invention is a kit comprising five poly-dT primers, wherein each of the primers comprises, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC. In specific embodiments the kit may further comprise a polymerase. Specifically the polymerase may be a reverse transcriptase or a DNA polymerase. It is contemplated that the kit further may comprise a label on the primers. Specifically it is contemplated that each primer may comprise a distinct label. The label may be a fluorometric label, colorimetric label, enzymatic label, biotin label, magnetic label or radioactive label. further, it is contemplated that the kit further comprises standard polynucleotides suitable for amplification by each of the primers. In the kit of the present it is contemplated that the poly-dT sequence is about 10 to about 35 bases. Specifically contemplated is a poly-dT sequence that is 11 bases. Of course this is only exemplary and sequences of 12 bases, 13 bases, 14 bases, 15 bases, 16 bases, 17 bases, 18 bases, 19 bases, 20 bases, 21 bases, 22 bases, 23 bases, 24 bases, 25 bases, 26 bases, 27 bases, 28 bases, 29 bases, 30 bases, 31 bases, 32 bases, 33 bases, 34 bases, 35 bases and longer also are specifically contemplated. Further it is contemplated that the kit may comprise arbitrary primers.

Also provided by the present invention is a method for the identifying an expressed gene fragment comprising (a) providing a polyA-minus cDNA population labeled at its 3'-end; (b) digesting the cDNA population with a restriction enzyme; (c) isolating the 3' fragments of the population; (d) performing 3' cDNA subtraction on

the fragments; (e) performing suppression PCR on the subtracted fragments; and (f) identifying a gene fragment from the amplified fragments.

In certain embodiments, it is contemplated that the method further comprises reverse transcribing an mRNA population into the cDNA population. In preferred embodiments the label is biotin. In particularly preferred embodiments, the primers used for reverse transcription consist of a poly-dT sequence and, at the 3' end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC. In specific embodiments, the primers further comprise the sequence TTTGCATGCTCGAG 5' to the poly-dT sequence. In specific embodiments, the poly-dT sequence is about 10 to about 35 bases. More specifically, the poly-dT sequence is 16 bases. In preferred embodiments, the restriction enzyme is *Nla*III. In other embodiments, the method further comprises verifying the subtraction efficiency. In specific embodiments, the verification is via multiplex quantitative PCR. In particularly defined embodiments, the targets for the PCR are one or more of the β-actin gene, the *HSC*70 gene and the *HSP*75 gene.

In certain aspects of the present invention, the method further may comprise cloning the isolated gene fragment. In additional embodiments the method further comprises sequencing of the cloned gene fragment. In specific embodiments, the sequencing is one-pass sequencing. In other particular embodiments the sequencing is SAGE sequencing. Particularly preferred embodiments comprise a method further comprising comparing the resulting sequence with one or more sequencing-containing databases. In especially preferred embodiments, the method further comprises identifying a plasmid containing the matched sequence from the I.M.A.G.E. consortium. In other preferred embodiments, the method further comprises probing a cDNA library with the cloned gene fragment. In more specific embodiments, the method further comprises isolating a complete cDNA corresponding to the cloned gene fragment. In additional embodiments, the method further comprises cloning the complete cDNA. In additional embodiments the method may comprise sequencing the cloned complete cDNA.

The invention also describes a method for the identifying an expressed gene fragment comprising: (a) converting mRNA molecules into a polydA/dT- minus

cDNA population; (b) digesting the cDNA population with a restriction enzyme; (c) isolating the 3' DNA fragments of the population thereby generating a 3' polydA/dT-minus cDNA library; (d) generating from the cDNA library: (i) a single-stranded cDNA library; and (ii) double-stranded inserts; (e) performing a subtraction on the single-stranded library using the double-stranded inserts; (f) eliminating double-stranded hybrids, thereby isolating a circular single-stranded cDNA sublibrary; and (g) sequencing the cDNA clones from step (f).

The method may further comprise prior to step (a) the step of obtaining mRNA molecules. The method also contemplates that the conversion of mRNA molecules into the cDNA population comprises the use of anchored dT primers, a polymerase and dNTPs. The anchored polydT primers are each comprised of a poly-dT sequence and, at the 3' end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of dA, dG, CA, CG and CC. In some aspects of the invention the poly-dT sequence is about 10 to about 35 bases. In a preferred aspect the poly-dT sequence is 16 bases.

In one preferred embodiment, the polymerase used is a reverse transcriptase. In more specific embodiments, the preferred reverse transcriptase is MMLV reverse transcriptase. In other specific embodiments, the reverse transcriptase is AMV reverse transcriptase. The preferred restriction enzyme used in the method is *Nla*III.

In one embodiment of the method, the generation of polydA/dT- minus 3' cDNA library comprises cloning the isolated 3' cDNA fragments. Several methods of sequencing are contemplated to be of use with the method as are known to one of skill in the art. In one aspect the sequencing is one-pass sequencing. In another aspect the sequencing is SAGE sequencing. In further aspects the method comprises comparing the sequence obtained with one or more sequence-containing databases.

Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and the specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various

changes and modifications within the spirit and scope of the invention will become apparent to those skilled in the art from this detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

The following drawings form part of the present specification and are included to further demonstrate certain aspects of the present invention. The invention may be better understood by reference to these drawings and the detailed description presented below.

FIG. 1. Schematic for Integrated Procedure for Gene Identification.

FIG. 2. Schematic for *In Vitro* Assay of RT of PolyA RNA Using Anchored Oligo-dT Primers.

FIG. 3. Sequence Pattern of MMLV RT generated by CT11. Color pattern: Green = A; Blue = G; Yellow = C; Red = T. The longer red strands are from the opposite cloning orientation. The correct cDNA products have only 11 A or T sequences from the primer itself. The misextended cDNA products have long polydA (green) or polydT (red) sequences.

FIG. 4. Hybridization Between PolydA/PolydT Sequences. This type of hybridization is not based on the existence of complementary strands, and will therefore lead to a heavy loss of low copy templates during the subtraction reaction.

FIG. 5. Potential Applications of the PolydA/PolydT minus cDNA Generated by Optimized Anchor Oligo-dT Primers.

FIG. 6. Size Distribution of *Nla*III Digested cDNA. Double-stranded cDNA was digested with *Nla*III and checked on an agarose gel to determine the range of size distribution.

FIG. 7. Determination of Subtraction Efficiency by Multiplex Quantitative PCR. The level of β-actin, Hsc70 and Hsp75 templates in different subtracted samples were quantified with the same set of control DNA and primers. The ratio

between each set of wild-type and control templates reflects their relative content in each sample.

FIGs. 8A-8D. Distribution of sequences collected in poly dA/dT(-) colon cDNA library. A total of 109 clones from un-normalized library and 193 clones from normalized library were sequenced. All the sequences were aligned with databases. SAGE: tags containing 10 nucleotides revere also collected from many of these clones and aligned with SAGE databases. **A.** Distribution of sequences from un-normalized library. **B.** Distribution of sequences from normalized library. **C.** Distribution of SAGE tags from un-normalized library. **D.** Distribution of SAGE tags from normalized library.

FIG. 9. Schematic for the SPGI method.

DETAILED DESCRIPTION OF THE INVENTION

To understand the gene expression pattern in cells under particular physiological and pathological conditions, the analysis must be performed at the genome scale. The EST project aims to collect expressed human sequences through screening many cDNA libraries from various sources (Boguski, 1995). Whereas the absolute number of EST sequences identified has steadily increased since the beginning of the EST project, the rate of novel gene identification from recently collected EST sequences is decreasing. This does not imply that most of the expressed human genes have been identified in EST sequences, however. A comparison between sequences from the EST project and from SAGE (Serial Analysis of Gene Expression), a technique for global gene identification (Velculescu *et al.*, 1995), shows that only 54% of SAGE sequences can be matched to existing sequences, including EST sequences (Zhang *et al.*, 1997).

One possibility for this observation is that the libraries used for the EST project failed to detect portions of the originally expressed population, particularly the low abundant sequences. Most cDNAs contain poly-dA/dT sequences at their 3′ ends, as they were generated by oligo-dT priming for reverse transcription (Bonaldo

et al., 1996). Before being used for sequencing, these libraries were normalized, or subtracted, in order to decrease the redundancy by removing high abundant copies (Bonaldo *et al.*, 1996). During these processes, nonspecific, randomized hybridization between poly-dA and poly-dT sequences among templates will occur in addition to the predicted formation of hybrids between the specific complementary templates, which leads to loss of many low abundant copies and, thus, results in the generation of incomplete libraries. This loss directly affects the efficiency of gene identification through the EST project. Therefore, it would be useful to generate poly-dA/dT minus cDNA populations for the analysis in order to avoid the loss of the low abundant copies.

Another important technique for analyzing large numbers of expressed nucleic acids is Differential Display. Almost every processed mRNA molecule contains between 50 to 250 bases of polyA sequences, which are important in controlling mRNA stability and metabolism (Karl, 1996; Colgan and Manley, 1997). These sequences provide an ideal target for oligo-dT-based primers in reverse transcription reactions. In 1992, Liang and Pardee first described Differential Display as a method of identifying differentially expressed genes at the genome level. In this technique, 3′-anchored oligo-dT primers (dT11VN, where V = random A, G, C and N=A, G, C, T) were used separately in reverse transcription reactions to generate cDNAs divided into twelve populations. Using these primers and 5′ 10-base arbitrary primers, the cDNAs were amplified and labeled using PCR™ and separated by denaturing gel electrophoresis. Comparison of the amplified products from different mRNA sources revealed variation in expression patterns in the corresponding cells. More than 1760 papers applying Differential Display can be found on the current Medline Database, demonstrating the popularity of this technique.

However, a general problem with this technique is the high false positive rate of gene identification. In most of the cases, only a limited number of real differentially expressed genes could be identified out of a large number of candidates. It is not uncommon that no candidate genes can be confirmed despite considerable effort. Various modifications have been implemented to improve Differential Display, focusing primarily on changing the length of the primers, the PCR temperatures, the blotting conditions, *etc.*. Another important modification is the use

of one-base-anchored oligo-dT primers (dT11A, dT11G, dT11C) for cDNA generation. However, there still remain serious question as to the fidelity of this system.

The subtraction methods, also have been utilized to examine expressed sequences. The subtraction technique is a powerful tool in gene identification analysis because it eliminates the abundant and intermediate abundant templates through a simple reassociation reaction, and enriches for differentially expressed genes (Duguid and Dinauer, 1990). However, major limitations of the classical subtraction methods are the need for a large quantity of initial material, and the difficulty in identifying the genes expressed at a less abundant level.

Another PCR™ subtraction based technique, called Suppression Subtraction Hybridization (SSH) has been recently developed (Diatchenko *et al.*, 1996). In this technique, the abundant and intermediate abundant templates present in two samples were eliminated through reassociation, and the differentially expressed templates were enriched through selective amplification *via* suppression PCR™. This technique requires much less mRNA than the classical subtraction technique, and can enrich some particular transcripts by a thousand-fold (Diatchenko *et al.*, 1996). However, the biggest limitation of this technique is the uncertainty of the sequence location within the gene due to lack of selection of specific regions of the gene. Thus, many sequences obtained from SSH cannot be matched to the EST sequences as they are not located in the 3′ or 5′ portion of the gene and these are incorrectly considered to be "novel sequences" representing potential novel genes (Diatchenko *et al.*, 1996). To clone these "novel" genes one would need to screen libraries, despite the fact that clones may already exist in EST plasmid stock.

Thus, in summary, though many powerful techniques exist to analyze expressed sequences in cells, each of these has significant limitations.

A. The Present Invention

1. Determination of Optimal Combination of Anchored Oligo-dT Primers for Reverse Transcription

The present invention impacts various of the preceding technologies by providing an improved set of anchored oligo-dT primers for use according to standard reverse transcription reactions. In order to provide fundamental information for the generation of poly-dA/dT minus cDNA populations with anchored primers, the inventors systematically studied the sequence patterns of cDNAs synthesized with anchored oligo-dT primers and reverse transcriptases. A correct product is extended from an anchored primer in which the anchoring nucleotide is correctly paired to the last nucleotide 5' to the poly A mRNA sequence. An erred product is extended from a primer with a non-paired anchoring nucleotide. It was determined, using single base 3'-anchored dT primers, that dT11C gave an error rate of 62% for MMLV reverse transcriptase and 72% for AMV reverse transcriptase. In contrast, dT11A and dT11G gave error rates of only about 3% for MMLV RT and 50% for AMV RT. Thus, the creation of variable length poly-dA sequences stemming from the use of C-anchored oligo-dT primers is the apparent reason for high false positive rate in the differential display technique.

In order to reduce the dT11C error rate, a second base was added to this primer, creating a family of primers: dT11CA, dT11CG, dT11CC and dT11CT. It was hoped that the additional base would increase the fidelity of priming and eliminate the spurious priming throughout the polyA sequence that give rise to variable length polyA cDNAs. As hoped, with the first three, the error rate dropped to 0%, 0% and about 33%, respectively. Surprisingly, the fourth, dT11CT, actually increased the error rate to about 84%. Thus, based on these studies, it is evident that the ideal set of primers is dT11A, dT11G, dT11CA, dT11CG and dT11CC. Using this combination, 91.75% of all possible sequences will be obtained. This primer set can be exploited without further modification in cDNA library construction from polyA messenger RNA populations, particularly useful for genome-wide gene identification through subtraction, and immediately applicable to replace the anchored primer set currently used in Differential Display.

2. The Establishment of an Integrated Procedure for Gene Identification: a Method for Genome-Wide Gene Identification

Using a new approach, rare abundant copies can be identified from a subtracted 3' cDNA population, and the EST database can be maximally used for

further analysis of these sequences, the rate at which bona fide novel sequences are identified can be increased. The procedure includes a) removing poly-dA sequences to avoid cross hybridization during the subtraction process; b) collecting only the 3′ portion cDNA from all expressed genes to insure that most of the sequences would be within the 3′ EST sequence range; c) performing subtraction hybridization to remove redundant templates, and selectively amplifying the enriched genes by suppression PCR™; d) verifying the subtraction efficiency by multiplex quantitative PCR™ instead of Northern blot to decrease the amount of mRNA required; and e) matching the sequences to databases like dbEST to identify the resultant sequences as existing genes, EST sequences or novel sequences, and to obtain additional information related to these sequences. The development of this method largely simplifies the process of genome-wide gene analysis. It can be used for any genome scale gene expression analysis, *e.g.*, EST project and the Cancer Genome Anatomy Project (CGAP).

Table 1 illustrates the IPGI approach with conventional methods. The embodiments and others are described more fully in the following pages.

Table 1

Original Techniques	Integrated Parts	Purpose
Differential Display	Anchored oligo dT primers for reverse transcription	Generate poly dA/polydT minus cDNAs to avoid the loss of low abundant copies in subtraction step
SAGE	NIaIII digestion Biotin labeling	Focus only on the 3' sequences of any gene to maximally identify genes through matching EST
SSH	Subtraction	Reduce mRNA requirement, remove the abundant copies
SSH	Suppressive PCR	Enrich the unsubtracted rare copies
Multiplex quantitative -PCR	Relative quantification	Determine the subtraction efficiency
EST/CGAP	Large scale sequencing	Index the expressed genes
SAGE	Sequencing only 14 bases for each template	Index the expressed gene in a much smaller scale Identify differentially expressed genes
GenBank database	NCBI blast search	Distinguish known genes, EST and novel sequences
IMAGE consortium	EST plasmid stock	Obtain clones containing the matched EST sequences

3. Screening Poly dA/dT(-) cDNAs For Gene Identification (SPGI) : another method for genome-wide gene identification.

The presence of long poly dA/dT sequences in the 3' end of cDNA templates contributes significantly to slowing the identification of novel genes from the human genome. The inventors developed a method, called screening poly dA/dT(-) cDNAs for gene identification (SPGI), and proved that it overcomes this obstacle and allows for the identification of high copy genes that can escape identification by methods known in the art due to the formation of double-stranded poly dA/dT hybrids and tangles between non-complementary genes. Applying this strategy to the generation of all the cDNA libraries enhances the efficiency of genome-wide gene identification, and thus will have major a impact on many functional genomic studies.

The SPGI method involves the following steps: a) converting mRNA molecules into a polydA/dT- minus cDNA population; b) digesting the cDNA population with a restriction enzyme; c) isolating the 3' fragments of the population thereby generating a 3' polydA/dT- minus cDNA library; e) creating from the polydA/dT- minus cDNA library a single-stranded cDNA library and double-stranded inserts; f) performing a subtraction on the single-stranded library using double-

stranded inserts; g) eliminating double-stranded hybrids thereby isolating a unique circular single-stranded cDNA sublibrary; and h) sequencing cDNA clones generated from the sublibrary for gene identification.

The inventors demonstrate in the examples that follow that the presence of poly dA/dT sequences in cDNA templates leads to the loss of cDNA templates upon subtraction. This loss contributes in large measure to the low efficiency of novel gene identification in the current EST/CGAP projects. This obstacle can be overcome through applying SPGI technique. The rate of gene identification in the current CGAP is about 4.6%, based on generation of 1,000 EST sequences per day (http://www.ncbi.nlm.nih.gov//ncicgap/). If one assume that there are about 30,000 unknown genes, and all of which would eventually be identified through the current EST/CGAP approaches then about 652,174 sequences will need to be identified in about 652 days. However, if the rate can be increased to 16% with SPGI strategy, the total sequencing effort can be decreased to 187,500 which could be completed in 187 days. This would be a significant increase in the efficiency of novel gene identification. In addition, the SPGI technique is also applicable in the functional genomic studies with various higher eukaryotic systems in the post-genome era.

B. Primers and Probes

1. Primer Design

The term primer, as defined herein, is meant to encompass any nucleic acid that is capable of priming the synthesis of a nascent nucleic acid in a template-dependent process. Typically, primers are oligonucleotides from ten to twenty-five base pairs in length, but longer sequences can be employed. Primers may be provided in double-stranded or single-stranded form, although the single-stranded form is preferred. Probes are defined differently, although they may act as primers. Probes, while perhaps capable of priming, are designed to binding to the target DNA or RNA and need not be used in an amplification process.

According to the present invention, there are disclosed, in one aspect, oligo-dT primers for use in reverse transcription and amplification reactions. These primers are 3′-anchored, *i.e.*, contain particular bases at their 3′ ends. These bases are the singlets A and G or the doublets, CC, CG or CA. This creates a set of five primers which give the highest possible coverage in random priming reactions (91.72%) without sacrifice of fidelity.

The particular length of the primer is not believed to be critical, with the dT sequence ranging from about 10 to about 25 bases, with 11 being a preferred embodiment. In some embodiments, the primers are labeled with radioactive species (^{32}P, ^{14}C, ^{35}S, ^{3}H, or other isotope), with a fluorophore (rhodamine, fluorescein, GFP) or a chemiluminescent label (luciferase).

Another type of primer, according to the present invention, is an arbitrary or random primer. Typically, such primers are used in combination with the anchored primer in a PCR-type reaction. The arbitrary primer serves to prime synthesis on the opposite strand as the anchored dT primer, permitting amplification. Such random primers are well known in the art and commercially available.

2. Probes

In various contexts, it may be useful to use oligo or polynucleotides as probes for complementary or hybridizing DNA or RNA molecules. In this regard, one may include particular "target" sequences in the oligos of the present invention in order to detect the products by probe hybridization. Alternatively, the probes may recognize unique sequences in the amplified regions upstream of the anchored oligo-dT primers.

3. Promoters

In certain embodiments of the present invention, the primers of the present invention may advantageously include sequences for promoters therein. For example, the T7, T3 or SP6 RNA polymerase promoters may be included in the primers used for amplification so that the resulting cDNA product includes one of these promoters, thereby permitting expression of an RNA transcript therefrom.

Another promoter suitable for inclusion in the primer constructs is the M13 phage promoter. This promoter permits rapid and fascile dideoxy sequencing of

cDNA after cloning into an M13-based vector. Examples of such vectors include pBluescript SK™ and pGEM32f(t).

4. Hybridization

Suitable hybridization conditions will be well known to those of skill in the art. Typically, the present invention relies on high stringency conditions (low salt, high temperature), which are well known in the art. Conditions may be rendered less stringent by increasing salt concentration and decreasing temperature. For example, a medium stringency condition could be provided by about 0.1 to 0.25 M NaCl at temperatures of about 37°C to about 55°C, while a low stringency condition could be provided by about 0.15 M to about 0.9 M salt, at temperatures ranging from about 20°C to about 55°C. Thus, hybridization conditions can be readily manipulated, and thus will generally be a method of choice depending on the desired results.

5. Primer Synthesis

Oligonucleotide synthesis is performed according to standard methods. See, for example, Itakura and Riggs (1980). Additionally, U. S. Patent No. 4,704,362; U. S. Patent No. 5,221,619 U. S. Patent No. 5,583,013 each describe various methods of preparing synthetic structural genes.

Oligonucleotide synthesis is well known to those of skill in the art. Various different mechanisms of oligonucleotide synthesis have been disclosed in for example, U.S. Patents. 4,659,774, 4,816,571, 5,141,813, 5,264,566, 4,959,463, 5,428,148, 5,554,744, 5,574,146, 5,602,244, each of which is incorporated herein by reference.

Basically, chemical synthesis can be achieved by the diester method, the triester method polynucleotides phosphorylase method and by solid-phase chemistry. These methods are discussed in further detail below.

Diester method. The diester method was the first to be developed to a usable state, primarily by Khorana and co-workers. (Khorana, 1979). The basic step is the joining of two suitably protected deoxynucleotides to form a dideoxynucleotide

containing a phosphodiester bond. The diester method is well established and has been used to synthesize DNA molecules (Khorana, 1979).

Triester method. The main difference between the diester and triester methods is the presence in the latter of an extra protecting group on the phosphate atoms of the reactants and products (Itakura *et al.*, 1975). The phosphate protecting group is usually a chlorophenyl group, which renders the nucleotides and polynucleotide intermediates soluble in organic solvents. Therefore purification's are done in chloroform solutions. Other improvements in the method include (i) the block coupling of trimers and larger oligomers, (ii) the extensive use of high-performance liquid chromatography for the purification of both intermediate and final products, and (iii) solid-phase synthesis.

Polynucleotide phosphorylase method. This is an enzymatic method of DNA synthesis that can be used to synthesize many useful oligodeoxynucleotides (Gillam *et al.*, 1978; Gillam *et al.*, 1979). Under controlled conditions, polynucleotide phosphorylase adds predominantly a single nucleotide to a short oligodeoxynucleotide. Chromatographic purification allows the desired single adduct to be obtained. At least a trimer is required to start the procedure, and this primer must be obtained by some other method. The polynucleotide phosphorylase method works and has the advantage that the procedures involved are familiar to most biochemists.

Solid-phase methods. Drawing on the technology developed for the solid-phase synthesis of polypeptides, it has been possible to attach the initial nucleotide to solid support material and proceed with the stepwise addition of nucleotides. All mixing and washing steps are simplified, and the procedure becomes amenable to automation. These syntheses are now routinely carried out using automatic DNA synthesizers.

Phosphoramidite chemistry (Beaucage and Lyer, 1992) has become by far the most widely used coupling chemistry for the synthesis of oligonucleotides. As is well known to those skilled in the art, phosphoramidite synthesis of oligonucleotides involves activation of nucleoside phosphoramidite monomer precursors by reaction with an activating agent to form activated intermediates, followed by sequential

addition of the activated intermediates to the growing oligonucleotide chain (generally anchored at one end to a suitable solid support) to form the oligonucleotide product.

C. Polymerases

1. Reverse Transcriptases

According to the present invention, a variety of different reverse transcriptases may be utilized. The following are representative examples.

M-MLV Reverse Transcriptase. M-MLV (Moloney Murine Leukemia Virus Reverse Transcriptase) is an RNA-dependent DNA polymerase requiring a DNA primer and an RNA template to synthesize a complementary DNA strand. The enzyme is a product of the *pol* gene of M-MLV and consists of a single subunit with a molecular weight of 71kDa. M-MLV RT has a weaker intrinsic RNase H activity than Avian Myeloblastosis Virus (AMV) reverse transcriptase which is important for achieving long full-length complementary DNA (>7 kB).

M-MLV can be use for first strand cDNA synthesis and primer extensions. Storage recommend at -20°C in 20 mM Tris-HCl (pH 7.5), 0.2M NaCl, 0.1 mM EDTA, 1 mM DTT, 0.01% Nonidet® P-40, 50% glycerol. The standard reaction conditions are 50 mM Tris-HCl (pH 8.3), 7 mM $MgCl_2$, 40 mM KCl, 10 mM DTT, 0.1 mg/ml BSA, 0.5 mM ^{3}H-dTTP, 0.025 mM oligo$(dT)_{50}$, 0.25 mM poly$(A)_{400}$ at 37°C.

M-MLV Reverse Transcriptase, RNase H Minus. This is a form of Moloney murine leukemia virus reverse transcriptase (RNA-dependent DNA polymerase) which has been genetically altered to remove the associated ribonuclease H activity (Tanese and Goff, 1988). It can be used for first strand cDNA synthesis and primer extension. Storage is at 20°C in 20 mM Tris-HCl (pH 7.5), 0.2M NaCl, 0.1 mM EDTA, 1 mM DTT, 0.01% Nonidet® P-40, 50% glycerol.

AMV Reverse Transcriptase. Avian Myeloblastosis Virus reverse transcriptase is a RNA dependent DNA polymerase that uses single-stranded RNA or DNA as a template to synthesize the complementary DNA strand (Houts *et al.*, 1979). It has activity at high temperature (42°C - 50°C). This polymerase has been used to synthesize long cDNA molecules.

Reaction conditions are 50 mM Tris-HCl (pH 8.3), 20 mM KCl, 10 mM $MgCl_2$, 500 μM of each dNTP, 5 mM dithiothreitol, 200 μg/ml oligo-$dT_{(12-18)}$, 250 μg/ml polyadenylated RNA, 6.0 pMol ^{32}P-dCTP, and 30 U enzyme in a 7 μl volume. Incubate 45 min at 42°C. Storage buffer is 200 mM KPO_4 (pH 7.4), 2 mM dithiothreitol, 0.2% Triton X-100, and 50% glycerol. AMV may be used for first strand cDNA synthesis, RNA or DNA dideoxy chain termination sequencing, and fill-ins or other DNA polymerization reactions for which Klenow polymerase is not satisfactory (Maniatis *et al.*, 1976).

2. DNA polymerases

The present invention also contemplates the use of various DNA polymerase. Exemplary polymerases are described below.

Bst DNA Polymerase, Large Fragment. *Bst* DNA Polymerase Large Fragment is the portion of the *Bacillus stearothermophilus* DNA Polymerase protein that contains the 5′→3′ polymerase activity, but lacks the 5′→3′ exonuclease domain. *BST* Polymerase Large Fragment is prepared from an *E. coli* strain containing a genetic fusion of the *Bacillus stearothermophilus* DNA Polymerase gene, lacking the 5′→3′ exonuclease domain, and the gene coding for *E. coli* maltose binding protein (MBP). The fusion protein is purified to near homogeneity and the MBP portion is cleaved off *in vitro*. The remaining polymerase is purified free of MBP (Iiyy *et al.*, 1991).

Bst DNA polymerase can be used in DNA sequencing through high GC regions (Hugh and Griffin, 1994; McClary et al., 1991) and Rapid Sequencing from nanogram amounts of DNA template (Mead et al., 1991). The reaction buffer is 1X ThermoPol Butter (20 mM Tris-HCl (pH 8.8 at 25°C), 10 mM KCl, 10 mM

$(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1% Triton X-100). Supplied with enzyme as a 10X concentrated stock.

Bst DNA Polymerase does not exhibit 3′→5′ exonuclease activity. 100 μ/ml BSA or 0.1% Triton X-100 is required for long term storage. Reaction temperatures above 70°C are not recommended. Heat inactivated by incubation at 80°C for 10 min. *Bst* DNA Polymerase cannot be used for thermal cycle sequencing. Unit assay conditions are 50 mM KCl, 20 mM Tris-HCl (pH 8.8), 10 mM $MgCl_2$, 30 nM M13mp18 ssDNA, 70 nM M13 sequencing primer (-47) 24 mer (NEB #1224), 200 μM daTP, 200 μM dCTP, 200 μM dGTP, 100 μM ^{3}H-dTTP, 100 μg/ml BSA and enzyme. Incubate at 65°C. Storage buffer is 50 mM KCl, 10 mM Tris-HCl (pH 7.5), 1 mM dithiothreitol, 0.1 mM EDTA, 0.1% Triton-X-100 and 50% glycerol. Storage is at -20°C.

VENT$_R$® DNA Polymerase and VENT$_R$® (exo$^-$) DNA Polymerase. Vent$_R$ DNA Polymerase is a high-fidelity thermophilic DNA polymerase. The fidelity of Vent$_R$ DNA Polymerase is 5-15-fold higher than that observed for Taq DNA Polymerase (Mattila *et al.*, 1991; Eckert and Kunkel, 1991). This high fidelity derives in part from an integral 3′→5′ proofreading exonuclease activity in Vent$_R$ DNA Polymerase (Mattila *et al.*, 1991; Kong *et al.*, 1993). Greater than 90% of the polymerase activity remains following a 1 h incubation at 95°C.

Vent$_R$ (exo-) DNA Polymerase has been genetically engineered to eliminate the 3′→5′ proofreading exonuclease activity associated with Vent$_R$ DNA Polymerase (Kong *et al.*, 1993). This is the preferred form for high-temperature dideoxy sequencing reactions and for high yield primer extension reactions. The fidelity of polymerization by this form is reduced to a level about 2-fold higher than that of Taq DNA Polymerase (Mattila *et al.*, 1991; Eckert and Kunkel, 1991). Vent$_R$ (exo-) DNA Polymerase is an excellent choice for DNA sequencing and is included in their CircumVent Sequencing Kit (see pages 118 and 121).

Both Vent$_R$ and Vent$_R$ (exo-) are purified from strains of *E. coli* that carry the Vent DNA Polymerase gene from the archaea *Thermococcus litoralis* (Perler *et al.*, 1992). The native organism is capable of growth at up to 98°C and was isolated from

a submarine thermal vent (Belkin and Jannasch, 1985). They are useful in primer extension, thermal cycle sequencing and high temperature dideoxy-sequencing.

DEEP VENT$_R$™ DNA Polymerase and DEEP VENT$_R$™ (exo$^-$) DNA Polymerase. Deep Vent$_R$ DNA Polymerase is the second high-fidelity thermophilic DNA polymerase available from New England Biolabs. The fidelity of Deep Vent$_R$ DNA Polymerase is derived in part from an integral 3′→5′ proofreading exonuclease activity. Deep Vent$_R$ is even more stable than Vent$_R$ at temperatures of 95 to 100°C (see graph).

Deep Vent$_R$ (exo-) DNA Polymerase has been genetically engineered to eliminate the 3′→5′ proofreading exonuclease activity associated with Deep Vent$_R$ DNA Polymerase. This exo- version can be used for DNA sequencing but requires different dNTP/ddNTP ratios than those used with Vent$_R$ (exo-) DNA Polymerase. Both Deep Vent$_R$ and Deep Vent$_R$ (exo-) are purified from a strain of *E. coli* that carries the Deep Vent$_R$ DNA Polymerase gene from *Pyrococcus species* GB-D (Perler *et al.*, 1996). The native organism was isolated from a submarine thermal vent at 2010 meters (Jannasch *et al.*, 1992) and is able to grow at temperatures as high as 104°C. Both enzymes can be used in primer extension, thermal cycle sequencing and high temperature dideoxy-sequencing.

T7 DNA Polymerase (unmodified). T7 DNA polymerase catalyzes the replication of T7 phage DNA during infection. The protein dimer has two catalytic activities: DNA polymerase activity and strong 3′→5′ exonuclease (Hori *et al.*, 1979; Engler *et al.*, 1983; Nordstrom *et al.*, 1981). The high fidelity and rapid extension rate of the enzyme make it particularly useful in copying long stretches of DNA template.

T7 DNA Polymerase consists of two subunits: T7 gene 5 protein (84 kilodaltons) and *E. coli* thioredoxin (12 kilodaltons) (Hori *et al.*, 1979; Studier *et al.*, 1990; Grippo and Richardson, 1971; Modrich and Richardson, 1975; Adler and Modrich, 1979). Each protein is cloned and overexpressed in a T7 expression system

in *E. coli* (Studier *et al.*, 1990). It can be used in second strand synthesis in site-directed mutagenesis protocols (Bebenek and Kunkel, 1989).

The reaction buffer is 1X T7 DNA Polymerase Buffer (20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM dithiothreitol). Supplement with 0.05 mg/ml BSA and dNTPs. Incubate at 37°C. The high polymerization rate of the enzyme makes long incubations unnecessary. T7 DNA Polymerase is not suitable for DNA sequencing.

Unit assay conditions are 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 1 mM dithiothreitol, 0.05 mg/ml BSA, 0.15 mM each dNTP, 0.5 mM heat denatured calf thymus DNA and enzyme. Storage conditions are 50 mM KPO_4 (pH 7.0), 0.1 mM EDTA, 1 mM dithiothreitol and 50% glycerol. Store at -20°C.

DNA Polymerase I (*E. coli*). DNA Polymerase I is a DNA-dependent DNA polymerase with inherent 3′→5′ and 5′→3′ exonuclease activities (Lehman, 1981). The 5′→3′ exonuclease activity removes nucleotides ahead of the growing DNA chain, allowing nick-translation. It is isolated from *E. coli* CM 5199, a lysogen carrying λ*polA* transducing phage (obtained from N.E. Murray) (Murray and Kelley, 1979). The phage in this strain was derived from the original *polA* phage encoding wild-type Polymerase I.

Applications include nick translation of DNA to obtain probes with a high specific activity (Meinkoth and Wahl, 1987) and second strand synthesis of cDNA (Gubler and Hoffmann, 1983; D'Alessio and Gerard, 1988). The reaction buffer is *E. coli* Polymerase I/Klenow Buffer (10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$, 7.5 mM dithiothreitol). Supplement with dNTPs.

DNase I is not included with this enzyme and must be added for nick translation reactions. Heat inactivation is for 20 min at 75°C. Unit assay conditions are 40 mM $KP0_4$ (pH 7.5), 6.6 mM $MgCl_2$, 1 mM 2-mercaptoethanol, 20 µM dAT copolymer, 33 µM dATP and 33 µM ^{3}H-dTTP. Storage conditions are 0.1 M $KP0_4$ (pH 6.5), 1 mM dithiothreitol, and 50% glycerol. Store at -20°C.

DNA Polymerase I, Large (Klenow) Fragment. Klenow fragment is a proteolytic product of *E. coli* DNA Polymerase I which retains polymerization and 3′→5′ exonuclease activity, but has lost 5′→3′ exonuclease activity. Klenow retains the polymerization fidelity of the holoenzyme without degrading 5′ termini.

A genetic fusion of the *E. coli polA* gene, that has its 5′→3′ exonuclease domain genetically replaced by maltose binding protein (MBP). Klenow Fragment is cleaved from the fusion and purified away from MBP. The resulting Klenow fragment has the identical amino and carboxy termini as the conventionally prepared Klenow fragment.

Applications include DNA sequencing by the Sanger dideoxy method (Sanger *et al.*, 1977), fill-in of 3′ recessed ends (Sambrook *et al.*, 1989), second-strand cDNA synthesis, random priming labeling and second strand synthesis in mutagenesis protocols (Gubler, 1987)

Reactions conditions are 1X *E. coli* Polymerase I/Klenow Buffer (10 mM Tris-HCl (pH 7.5), 5 mM MgCl2, 7.5 mM dithiothreitol). Supplement with dNTPs (not included). Klenow fragment is also 50% active in all four standard NEBuffers when supplemented with dNTPs. Heat inactivated by incubating at 75°C for 20 min. Fill-in conditions: DNA should be dissolved, at a concentration of 50 μg/ml, in one of the four standard NEBuffers (1X) supplemented with 33 μM each dNTP. Add 1 unit Klenow per μg DNA and incubate 15 min at 25°C. Stop reaction by adding EDTA to 10 mM final concentration and heating at 75°C for 10 min. Unit assay conditions 40 mM KP04 (pH 7.5), 6.6 mM MgCl2, 1 mM 2-mercaptoethanol, 20 μM dAT copolymer, 33 μM dATP and 33 μM ^{3}H-dTTP. Storage conditions are 0.1 M $KP0_4$ (pH 6.5), 1 mM dithiothreitol, and 50% glycerol. Store at -20°C.

Klenow Fragment (3′→5′ exo⁻). Klenow Fragment (3′→5′ exo-) is a proteolytic product of DNA Polymerase I which retains polymerase activity, but has a mutation which abolishes the 3′→5′ exonuclease activity and has lost the 5′→3′ exonuclease (Derbyshire *et al.*, 1988).

A genetic fusion of the *E. coli polA* gene, that has its 3′→5′ exonuclease domain genetically altered and 5′→3′ exonuclease domain replaced by maltose binding protein (MBP). Klenow Fragment exo- is cleaved from the fusion and purified away from MBP. Applications include random priming labeling, DNA sequence by Sanger dideoxy method (Sanger *et al.*, 1977), second strand cDNA synthesis and second strand synthesis in mutagenesis protocols (Gubler, 1987).

Reaction buffer is 1X *E. coli* Polymerase I/Klenow Buffer (10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$, 7.5 mM dithiothreitol). Supplement with dNTPs. Klenow Fragment exo- is also 50% active in all four standard NEBuffers when supplemented with dNTPs. Heat inactivated by incubating at 75°C for 20 min. When using Klenow Fragment (3′→5′ exo-) for sequencing DNA using the dideoxy method of Sanger *et al.* (1977), an enzyme concentration of 1 unit/5 µl is recommended.

Unit assay conditions are 40 mM $KP0_4$ (pH 7.5), 6.6 mM $MgCl_2$, 1 mM 2-mercaptoethanol, 20 µM dAT copolymer, 33 µM dATP and 33 µM ^{3}H-dTTP. Storage conditions are 0.1 M $KP0_4$ (pH 7.5), 1 mM dithiothreitol, and 50% glycerol. Store at -20°C.

T4 DNA Polymerase. T4 DNA Polymerase catalyzes the synthesis of DNA in the 5′→3′ direction and requires the presence of template and primer. This enzyme has a 3′→5′ exonuclease activity which is much more active than that found in DNA Polymerase I. Unlike *E. coli* DNA Polymerase I, T4 DNA Polymerase does not have a 5′→3′ exonuclease function.

Purified from a strain of *E. coli* that carries a T4 DNA Polymerase overproducing plasmid. Applications include removing 3' overhangs to form blunt ends (Tabor and Struhl, 1989; Sambrook *et al.*, 1989), 5' overhang fill-in to form blunt ends (Tabor and Struhl, 1989; Sambrook *et al.*, 1989), single strand deletion subcloning (Dale *et al.*, 1985), second strand synthesis in site-directed mutagenesis (Kunkel *et al.*, 1987), and probe labeling using replacement synthesis (Tabor and Struhl, 1989; Sambrook *et al.*, 1989).

The reaction buffer is 1X T4 DNA Polymerase Buffer (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM dithiothreitol (pH 7.9 at 25°C)). Supplement with 40 μg/ml BSA and dNTPs (not included in supplied 10X buffer). Incubate at temperature suggested for specific protocol.

It is recommended to use 100 μM of each dNTP, 1-3 units polymerase/μg DNA and incubation at 12°C for 20 min in the above reaction buffer (Tabor and Struhl, 1989; Sambrook *et al.*, 1989). Heat inactivated by incubating at 75°C for 10 min. T4 DNA Polymerase is active in all four standard NEBuffers when supplemented with dNTPs.

Unit assay conditions are 50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM dithiothreitol (pH 7.9 at 25°C), 33 μM dATP, dCTP and dGTP, 33 μM ^3H dTTP, 70 μg/ml denatured calf thymus DNA, and 170 μg/ml BSA. Note: These are not suggested reaction conditions; refer to Reaction Buffer. Storage conditions are 100 mM KPO_4 (pH 6.5), 10 mM 2-mercaptoethanol and 50% glycerol. Store at -20°C.

3. RNA polymerases

RNA polymerases for use in the present invention are exemplified as follows.

T7 RNA Polymerase SP6 RNA Polymerase and T3 RNA Polymerase. Initiation of transcription with T7, SP6 RNA and **T3 RNA Polymerase** Polymerases is highly specific for the T7 and SP6 phage promoters, respectively. Cloning vectors have been developed which direct transcription from the T7 SP6 or T3 promoter through polylinker cloning sites (Schenborn and Meirendorf, 1985). These vectors allow *in vitro* synthesis of defined RNA transcripts from a cloned DNA sequence. Under optimal conditions, greater than 700 moles of T7 RNA transcript can be synthesized per mole of DNA template (Noren *et al.*, 1990). RNA produced using the SP6 and T7 RNA polymerases is biologically active as mRNA (Krieg and Melton, 1984) and can be accurately spliced (Green *et al.*, 1983). Anti-sense RNA, produced by reversing the orientation of the cloned DNA insert, has been shown to specifically block mRNA translation *in vivo* (Melton, 1985).

Labeled single-stranded RNA transcripts of high specific activity are simple to prepare with T7 and SP6 RNA polymerases (Sambrook *et al.*, 1989). Increased levels of detection in nucleic acid hybridization reactions can also be obtained due to the greater stability of RNA:DNA hybrids with respect to RNA:RNA or DNA:DNA hybrids (Zinn *et al.*, 1983).

SP6 RNA Polymerase is isolated form SP6 phage-infected *Salmonella typhimurium* LT2Z (Butler and Chamberlin, 1982). T7 RNA Polymerase is isolated from *E. coli* BL21 carrying the plasmid pAR1219 which contains T7 *gene l* under the control of the inducible *lac* UV6 promoter (Davanloo *et al.*, 1984). Applications include preparation of radiolabeled RNA probes (Sambrook *et al.*, 1989), RNA generation for *in vitro* translation (Sambrook *et al.*, 1989), RNA generation for studies of RNA structure, processing and catalysis (Sambrook *et al.*, 1989) and expression control via antisense RNA.

Reaction 1X RNA Polymerase Buffer: (40 mM Tris-HCl (pH 7.9), 6 mM $MgCl_2$, 2 mM spermidine, 10 mM dithiothreitol). Supplement with 0.5 mM each ATP, UTP, GTP, CTP (not included) and DNA template containing the appropriate promoter. Incubate at 37°C (T7 RNA polymerase) or 40°C (SP6 RNA polymerase).

Dithiothreitol is required for activity. Both enzymes are extremely sensitive to salt inhibition. For best results overall salt concentration should not exceed 50 mM. SP6 RNA polymerase is 30% more active at 40°C than at 37°C. Higher yields of RNA may be obtained by raising NTP concentrations (up to 4 mM each). Mg^{2+} concentration should be raised to 4 mM above the total NTP concentration. Additionally, inorganic pyrophosphatase should be added to a final concentration of 4 units/ml. SP6 RNA polymerase is supplied with a control template (NEB#207B). The template is a pSP64 vector containing a 1.38 kB insert, linearized at 3 different restriction sites. Transcription with SP6 RNA polymerase results in three runoff fragments of 1.38 kB, 0.55 kB and 0.22 kB.

Storage conditions are 100 mM NaCl, 50 mM Tris-HCl (pH 7.9), 1 mM EDTA, 20 mM 2-mercaptoethanol, 0.1% Triton-X-100 and 50% glycerol. Store at -20°C.

T3 RNA polymerase is a DNA dependent RNA polymerase which exhibits extremely high specificity for T3 promoter sequences. The enzyme will incorporates 32P, 35S and 3H-labeled nucleotide triphosphates. It is used in the synthesis of RNA transcripts for hybridization probes in vitro translation, RNase protection assays or RNA processing substrates.

One unit of T3 RNA polymerase is defined as the amount of enzyme required to catalyze the incorporation of 5nmol of CTP into acid insoluble product in 60 minutes at 37°C in a total volume of 100µl. The reaction conditions are as follows, 40mM Tris-HCl (pH 7.9), 6 mM $MgCl_2$, 10mM DTT, 10mM NaCl, 2mM spermidine, 0.5% Tween®-20, 0.5mM each ATP, GTP, DTP, and UTP, 0.5µCi [^{3}H] CTP, and 2µg supercoiled pSP6/T3 Vector DNA. Promega provide a T3 RNA polymerase extracted from recombinant *E. coli.*

D. Amplification Methodologies

A number of template dependent processes are available to amplify the marker sequences present in a given template sample. One of the best known amplification methods is the polymerase chain reaction (referred to as PCR™) which is described in detail in U.S. Patent Nos. 4,683,195, 4,683,202 and 4,800,159, and in Innis *et al.*, 1990.

Briefly, in PCR, two primer sequences are prepared that are complementary to regions on opposite complementary strands of the marker sequence. An excess of deoxynucleoside triphosphates are added to a reaction mixture along with a DNA polymerase, *e.g.*, *Taq* polymerase. If the marker sequence is present in a sample, the primers will bind to the marker and the polymerase will cause the primers to be extended along the marker sequence by adding on nucleotides. By raising and lowering the temperature of the reaction mixture, the extended primers will dissociate from the marker to form reaction products, excess primers will bind to the marker and to the reaction products and the process is repeated.

A reverse transcriptase PCR amplification procedure may be performed in order to amplify mRNA templates. Methods of reverse transcribing RNA into cDNA are well known and described in Sambrook *et al.*, 1989. Alternative methods for

reverse transcription utilize thermostable, RNA-dependent DNA polymerases. These methods are described in WO 90/07641 filed December 21, 1990. Polymerase chain reaction methodologies are well known in the art.

Another method for amplification is the ligase chain reaction ("LCR"), disclosed in EP No. 320 308. In LCR, two complementary probe pairs are prepared, and in the presence of the target sequence, each pair will bind to opposite complementary strands of the target such that they abut. In the presence of a ligase, the two probe pairs will link to form a single unit. By temperature cycling, as in PCR, bound ligated units dissociate from the target and then serve as "target sequences" for ligation of excess probe pairs. U.S. Patent 4,883,750 describes a method similar to LCR for binding probe pairs to a target sequence.

Qbeta Replicase, described in PCT Application No. PCT/US87/00880, may also be used as still another amplification method in the present invention. In this method, a replicative sequence of RNA that has a region complementary to that of a target is added to a sample in the presence of an RNA polymerase. The polymerase will copy the replicative sequence that can then be detected.

An isothermal amplification method, in which restriction endonucleases and ligases are used to achieve the amplification of target molecules that contain nucleotide 5'-[alpha-thio]-triphosphates in one strand of a restriction site may also be useful in the amplification of nucleic acids in the present invention, Walker *et al.* (1992).

Strand Displacement Amplification (SDA) is another method of carrying out isothermal amplification of nucleic acids which involves multiple rounds of strand displacement and synthesis, *i.e.*, nick translation. A similar method, called Repair Chain Reaction (RCR), involves annealing several probes throughout a region targeted for amplification, followed by a repair reaction in which only two of the four bases are present. The other two bases can be added as biotinylated derivatives for easy detection. A similar approach is used in SDA. Target specific sequences can also be detected using a cyclic probe reaction (CPR). In CPR, a probe having 3' and 5' sequences of non-specific DNA and a middle sequence of specific RNA is hybridized to DNA that is present in a sample. Upon hybridization, the reaction is

treated with RNase H, and the products of the probe identified as distinctive products that are released after digestion. The original template is annealed to another cycling probe and the reaction is repeated.

Still another amplification methods described in GB Application No. 2 202 328, and in PCT Application No. PCT/US89/01025, may be used in accordance with the present invention. In the former application, "modified" primers are used in a PCR-like, template- and enzyme-dependent synthesis. The primers may be modified by labeling with a capture moiety (*e.g.*, biotin) and/or a detector moiety (*e.g.*, enzyme). In the latter application, an excess of labeled probes are added to a sample. In the presence of the target sequence, the probe binds and is cleaved catalytically. After cleavage, the target sequence is released intact to be bound by excess probe. Cleavage of the labeled probe signals the presence of the target sequence.

Other nucleic acid amplification procedures include transcription-based amplification systems (TAS), including nucleic acid sequence based amplification (NASBA) and 3SR (Kwoh *et al.*, 1989; Gingeras *et al.*, PCT Application WO 88/10315). In NASBA, the nucleic acids can be prepared for amplification by standard phenol/chloroform extraction, heat denaturation of a clinical sample, treatment with lysis buffer and minispin columns for isolation of DNA and RNA or guanidinium chloride extraction of RNA. These amplification techniques involve annealing a primer which has target specific sequences. Following polymerization, DNA/RNA hybrids are digested with RNase H while double-stranded DNA molecules are heat denatured again. In either case the single stranded DNA is made fully double-stranded by addition of second target specific primer, followed by polymerization. The double-stranded DNA molecules are then multiply transcribed by an RNA polymerase such as T7 or SP6. In an isothermal cyclic reaction, the RNA's are reverse transcribed into single stranded DNA, which is then converted to double-stranded DNA, and then transcribed once again with an RNA polymerase such as T7 or SP6. The resulting products, whether truncated or complete, indicate target specific sequences.

Davey *et al.*, EP No. 329 822 disclose a nucleic acid amplification process involving cyclically synthesizing single-stranded RNA ("ssRNA"), ssDNA, and

double-stranded DNA (dsDNA), which may be used in accordance with the present invention. The ssRNA is a template for a first primer oligonucleotide, which is elongated by reverse transcriptase (RNA-dependent DNA polymerase). The RNA is then removed from the resulting DNA:RNA duplex by the action of ribonuclease H (RNase H, an RNase specific for RNA in duplex with either DNA or RNA). The resultant ssDNA is a template for a second primer, which also includes the sequences of an RNA polymerase promoter (exemplified by T7 RNA polymerase) 5' to its homology to the template. This primer is then extended by DNA polymerase (exemplified by the large "Klenow" fragment of *E. coli* DNA polymerase I), resulting in a double-stranded DNA ("dsDNA") molecule, having a sequence identical to that of the original RNA between the primers and having additionally, at one end, a promoter sequence. This promoter sequence can be used by the appropriate RNA polymerase to make many RNA copies of the DNA. These copies can then re-enter the cycle leading to very swift amplification. With proper choice of enzymes, this amplification can be done isothermally without addition of enzymes at each cycle. Because of the cyclical nature of this process, the starting sequence can be chosen to be in the form of either DNA or RNA.

Miller *et al.*, PCT Application WO 89/06700 disclose a nucleic acid sequence amplification scheme based on the hybridization of a promoter/primer sequence to a target single-stranded DNA ("ssDNA") followed by transcription of many RNA copies of the sequence. This scheme is not cyclic, *i.e.*, new templates are not produced from the resultant RNA transcripts. Other amplification methods include "RACE" and "one-sided PCR" (Frohman, M.A., In: *PCR PROTOCOLS: A GUIDE TO METHODS AND APPLICATIONS*, Academic Press, N.Y., 1990; Ohara *et al.*, 1989).

Methods based on ligation of two (or more) oligonucleotides in the presence of nucleic acid having the sequence of the resulting "di-oligonucleotide", thereby amplifying the di-oligonucleotide, may also be used in the amplification step of the present invention. Wu *et al.* (1989).

E. Differential Display

RNA fingerprinting is a means by which RNAs isolated from many different tissues, cell types or treatment groups may be sampled simultaneously to identify RNAs whose relative abundances vary. Two forms of this technology were developed simultaneously and reported in 1992 as RNA fingerprinting and differential display (Liang and Pardee, 1992; Welsh *et al.*, 1992). (*See also* Liang and Pardee, U.S. Patent 5,262,311, U.S. Patent 5,665,547 incorporated herein by reference in its entirety.) Both techniques were utilized in the studies described below. Some of the studies described herein were performed similarly to Donahue *et al.*, 1994.

All forms of RNA fingerprinting by PCR are theoretically similar but differ in their primer design and application. The most striking difference between differential display and other methods of RNA fingerprinting is that differential display utilizes anchoring primers that hybridize to the polyA tails of mRNAs. As a consequence, the PCR products amplified in differential display are biased towards the 3' untranslated regions of mRNAs.

The basic technique of differential display has been described in detail (Liang and Pardee, 1992). Total cell RNA is primed for first strand reverse transcription with an anchored primer composed of oligo-dT. The oligo-dT primer is extended using a reverse transcriptase, for example, Moloney Murine Leukemia Virus (MMLV) reverse transcriptase. The synthesis of the second strand is primed with an arbitrarily chosen oligonucleotide, using reduced stringency conditions. Once the double-stranded cDNA has been synthesized, amplification proceeds by standard PCR techniques, utilizing the same primers. The resulting DNA fingerprint is analyzed by gel electrophoresis with ethidium bromide staining or autoradiography. A side by side comparison of fingerprints from different cell derived RNAs using the same oligonucleotide primers identifies mRNAs that are differentially expressed.

Differential display technology has been demonstrated as being effective in identifying genes that are differentially expressed in cancer (Liang and Pardee, 1992; Wong *et al.*, 1993; Sager *et al.*, 1993; Mok *et al.*, 1994; Watson *et al.*, 1994; Chen *et al.*,

1995; An *et al.*, 1995). The present invention utilizes the RNA fingerprinting technique to identify genes that are differentially expressed in prostate cancer. These studies utilized RNAs isolated from tumor tissues and tumor-derived cell lines that behave as tumors cells with different metastatic potential.

The underlying concept of these studies was that genes that are differentially expressed in cells with different metastatic potentials may be used as indicators of metastatic potential. Since metastasis is a prerequisite for prostate cancer progression to life threatening pathologies, indicators of metastatic potential are likely to be indicators of pathological potential.

Cells often are harvested in late log phase of growth. RNA may be isolated by the guanidinium thiocyanate method (Chomczynski and Sacchi, 1987). After RNA isolation, the nucleic acids are precipitated with ethanol. The precipitates are pelleted by centrifugation and redissolved in water. The redissolved nucleic acids are then digested with RNase-free DNase I (Boehringer Mannheim, Inc.) following the manufacturer's instructions, followed by organic extraction with phenol:chloroform:isoamyl alcohol (25:24:1) and reprecipitation with ethanol.

The DNase I treated RNA is then pelleted by centrifugation and redissolved in water. The purity and concentration of the RNA in solution is estimated by determining optical density at wave lengths of 260 nm and 280 nm (Sambrook *et al.*, 1989). A small aliquot of the RNA is separated by gel electrophoresis in a 3% formaldehyde gel with MOPS buffer (Sambrook *et al.*, 1989) to confirm the estimation of concentration and to determine if the ribosomal RNAs were intact. This RNA is referred to as total cell RNA.

There were two kinds of RNA fingerprinting studies performed with the total cell RNA. The first of these kinds of studies follow the differential display protocol of Liang and Pardee (1992) except that they are modified by using 5' biotinylated primers for nonisotopic PCR product detection.

In these studies, 0.2 μg of total cell RNA are primed for reverse transcription with an anchoring primer according to the present invention, then two arbitrarily chosen

nucleotides, including all of the possible combinations of each nucleotide at these positions. Reverse transcription is performed with 200 units of MMLV (Moloney Murine Leukemia Virus) reverse transcriptase (GIBCO/BRL) in the presence of 50 mM Tris-HCl (pH 8.3), 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 500 μM dNTP, 1 μM anchored primer and 1 U/μl RNase inhibitor. The reaction mixture is incubated at room temperature for 10 minutes, then at 37°C for 50 minutes. After reverse transcription the enzyme is inactivated by heating to 65°C for 10 minutes.

One tenth of the resulting reverse transcription reactions is then amplified by PCR using the same anchoring primer as used in the reverse transcription step and a second oligonucleotide of arbitrarily chosen sequences. The PCR reaction contains 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 20 μM dNTP, 1.5 μM $MgCl_2$, 200 nM arbitrary decamer, 1 μM anchored primer, and 1 unit of Taq DNA polymerase (Boehringer Mannheim) in a 40 μl volume. The amplification is performed in a thermal cycler (MJ Research) for 30 cycles with denaturing at 94°C for 30 sec, annealing at 40°C for 2 min, and extending at 72°C for 30 sec, ^{35}S-dATP is added in the PCR reaction.

The PCR products are then separated on a 6% TBE-urea sequencing gel (Sambrook *et al.*, 1989) and detected by autoradiography. Differentially appearing PCR products may be excised from the gels, reamplified using the same primers used in the original amplification, and cloned using the TA cloning strategy (Invitrogen, Inc. and Promega, Inc.).

The second type of RNA fingerprinting studies more closely resembled the protocol of Welsh *et al.* (1992). This approach uses a variation of the above as modified by the use of agarose gels and non-isotopic detection of bands by ethidium bromide staining (An *et al.*, 1995). Total RNAs are isolated from the frozen prostate tissues or cultured cells as described (Chomczynski and Sacchi, 1987). Ten micrograms of total cellular RNAs are treated with 5 units of RNAse-free DNAse I (GIBCO/BRL) in 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2 mM $MgCl_2$, and 20 units of RNAse inhibitor (Boehringer Mannheim). After extraction with phenol/chloroform and ethanol precipitation, the RNAs are redissolved in DEPC-treated water.

Two μg of each total cell RNA sample are reverse transcribed into cDNA using randomly selected hexamer primers and MMLV reverse transcriptase (GIBCO/BRL). PCR was performed using one or two arbitrarily chosen oligonucleotide primers (10-12mers). PCR conditions are: 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 50 μM dNTPs, 0.2 μM of primer(s), 1 unit of Taq DNA polymerase (GIBCO/BRL) in a final volume of 20 μl. The amplification parameters include 35 cycles of reaction with 30 sec denaturing at 94°C, 90 sec annealing at 40°C, and 60 sec extension at 72°C. A final extension at 72°C is performed for 15 min. The resulting PCR products are resolved into a fingerprint by size separation by electrophoresis through 2% agarose gels in TBE buffer (Sambrook *et al.*, 1989). The fingerprints are visualized by staining with ethidium bromide. No reamplification is performed.

Differentially appearing PCR products, that might represent differentially expressed genes, are excised from the gel with a razor blade, purified from the agarose using the Geneclean kit (Bio 101, Inc.), eluted in water and cloned directly into plasmid vectors using the TA cloning strategy (Invitrogen, Inc., and Promega, Inc.). These products are not reamplified after the initial PCR fingerprinting protocol.

F. Blotting Techniques

Blotting techniques are well known to those of skill in the art. Southern blotting involves the use of DNA as a target, whereas Northern blotting involves the use of RNA as a target. Each provide different types of information, although cDNA blotting is analogous, in many aspects, to blotting or RNA species.

Briefly, a probe is used to target a DNA or RNA species that has been immobilized on a suitable matrix, often a filter of nitrocellulose. The different species should be spatially separated to facilitate analysis. This often is accomplished by gel electrophoresis of nucleic acid species followed by "blotting" on to the filter.

Subsequently, the blotted target is incubated with a probe (usually labeled) under conditions that promote denaturation and rehybridization. Because the probe is designed to base pair with the target, the probe will bind a portion of the target

sequence under renaturing conditions. Unbound probe is then removed, and detection is accomplished as described above.

G. Separation Methods

It normally is desirable, at one stage or another, to separate the amplification products from reagents, such as the template or excess primers, or from other amplification products. In one embodiment, amplification products are separated by agarose, agarose-acrylamide or polyacrylamide gel electrophoresis using standard methods. See Sambrook *et al.*, 1989. When working with nucleic acids, denaturing PAGE is preferred.

Alternatively, chromatographic techniques may be employed to effect separation. There are many kinds of chromatography which may be used in the present invention: adsorption, partition, ion-exchange and molecular sieve, and many specialized techniques for using them including column, paper, thin-layer and gas chromatography (Freifelder, 1982).

H. Expression Vectors

Within certain embodiments, expression vectors are employed to express various polynucleotides in accordance with the present invention. Expression requires that appropriate signals be provided in the vectors, which include various regulatory elements, such as enhancers/promoters from both viral and mammalian sources that drive expression of the genes of interest in host cells. Elements designed to optimize messenger RNA stability and translatability in host cells also are defined. The conditions for the use of a number of dominant drug selection markers for establishing permanent, stable cell clones expressing the products are also provided, as is an element that links expression of the drug selection markers to expression of the polypeptide.

1. Regulatory Elements

Throughout this application, the term "expression construct" is meant to include any type of genetic construct containing a nucleic acid coding for a gene product in which part or all of the nucleic acid encoding sequence is capable of being

transcribed. The transcript may be translated into a protein, but it need not be. In certain embodiments, expression includes both transcription of a gene and translation of mRNA into a gene product. In other embodiments, expression only includes transcription of the nucleic acid encoding a gene of interest.

In preferred embodiments, the nucleic acid encoding a gene product is under transcriptional control of a promoter. A "promoter" refers to a DNA sequence recognized by the synthetic machinery of the cell, or introduced synthetic machinery, required to initiate the specific transcription of a gene. The phrase "under transcriptional control" means that the promoter is in the correct location and orientation in relation to the nucleic acid to control RNA polymerase initiation and expression of the gene.

The term promoter will be used here to refer to a group of transcriptional control modules that are clustered around the initiation site for RNA polymerase II. Much of the thinking about how promoters are organized derives from analyses of several viral promoters, including those for the HSV thymidine kinase (*tk*) and SV40 early transcription units. These studies, augmented by more recent work, have shown that promoters are composed of discrete functional modules, each consisting of approximately 7-20 bp of DNA, and containing one or more recognition sites for transcriptional activator or repressor proteins.

At least one module in each promoter functions to position the start site for RNA synthesis. The best known example of this is the TATA box, but in some promoters lacking a TATA box, such as the promoter for the mammalian terminal deoxynucleotidyl transferase gene and the promoter for the SV40 late genes, a discrete element overlying the start site itself helps to fix the place of initiation.

Additional promoter elements regulate the frequency of transcriptional initiation. Typically, these are located in the region 30-110 bp upstream of the start site, although a number of promoters have recently been shown to contain functional elements downstream of the start site as well. The spacing between promoter elements frequently is flexible, so that promoter function is preserved when elements are inverted or moved relative to one another. In the tk promoter, the spacing between promoter elements can be increased to 50 bp apart before activity begins to

decline. Depending on the promoter, it appears that individual elements can function either co-operatively or independently to activate transcription.

The particular promoter employed to control the expression of a nucleic acid sequence of interest is not believed to be important, so long as it is capable of direction the expression of the nucleic acid in the targeted cell. Thus, where a human cell is targeted, it is preferable to position the nucleic acid coding region adjacent to and under the control of a promoter that is capable of being expressed in a human cell. Generally speaking, such a promoter might include either a human or viral promoter.

In various embodiments, the human cytomegalovirus (CMV) immediate early gene promoter, the SV40 early promoter, the Rous sarcoma virus long terminal repeat, rat insulin promoter and glyceraldehyde-3-phosphate dehydrogenase can be used to obtain high-level expression of the coding sequence of interest. The use of other viral or mammalian cellular or bacterial phage promoters which are well-known in the art to achieve expression of a coding sequence of interest is contemplated as well, provided that the levels of expression are sufficient for a given purpose.

By employing a promoter with well-known properties, the level and pattern of expression of the protein of interest following transfection or transformation can be optimized. Further, selection of a promoter that is regulated in response to specific physiologic signals can permit inducible expression of the gene product.

Enhancers are genetic elements that increase transcription from a promoter located at a distant position on the same molecule of DNA. Enhancers are organized much like promoters. That is, they are composed of many individual elements, each of which binds to one or more transcriptional proteins.

The basic distinction between enhancers and promoters is operational. An enhancer region as a whole must be able to stimulate transcription at a distance; this need not be true of a promoter region or its component elements. On the other hand, a promoter must have one or more elements that direct initiation of RNA synthesis at a particular site and in a particular orientation, whereas enhancers lack these specificities. Promoters and enhancers are often overlapping and contiguous, often seeming to have a very similar modular organization.

Where a cDNA insert is employed, one will typically desire to include a polyadenylation signal to effect proper polyadenylation of the gene transcript. The nature of the polyadenylation signal is not believed to be crucial to the successful practice of the invention, and any such sequence may be employed such as human growth hormone and SV40 polyadenylation signals. Also contemplated as an element of the expression cassette is a terminator. These elements can serve to enhance message levels and to minimize read through from the cassette into other sequences.

2. Selectable Markers

In certain embodiments of the invention, the cells contain nucleic acid constructs of the present invention, a cell may be identified *in vitro* or *in vivo* by including a marker in the expression construct. Such markers would confer an identifiable change to the cell permitting easy identification of cells containing the expression construct. Usually the inclusion of a drug selection marker aids in cloning and in the selection of transformants, for example, genes that confer resistance to neomycin, puromycin, hygromycin, DHFR, GPT, zeocin and histidinol are useful selectable markers. Alternatively, enzymes such as herpes simplex virus thymidine kinase (*tk*) or chloramphenicol acetyltransferase (CAT) may be employed. Immunologic markers also can be employed. The selectable marker employed is not believed to be important, so long as it is capable of being expressed simultaneously with the nucleic acid encoding a gene product. Further examples of selectable markers are well known to one of skill in the art.

I. Kits

All the essential materials and reagents required for performing Differential Display, creating cDNA libraries and performing Suppression Subtraction Hybridization may be assembled together in a kit. Such kits generally will comprise preselected primers and may include other oligo- and polynucleotides, such as probes and expression vectors. Also included may be enzymes suitable for amplifying nucleic acids including various polymerases (reverse transcriptases, Taq, Sequenase™, *etc.*), dNTPs and buffers to provide the necessary reaction mixture for amplification. Such kits also generally will comprise, in suitable means, distinct containers for each individual primer, probe, vector, dNTPs, buffer and enzyme.

J. Examples

The following examples are included to demonstrate preferred embodiments of the invention. It should be appreciated by those of skill in the art that the techniques disclosed in the examples which follow represent techniques discovered by the inventor to function well in the practice of the invention, and thus can be considered to constitute preferred modes for its practice. However, those of skill in the art should, in light of the present disclosure, appreciate that many changes can be made in the specific embodiments which are disclosed and still obtain a like or similar result without departing from the spirit and scope of the invention.

Example 1: Materials and Methods

Oligo synthesis. The single-strand DNA template, PCR primers, and anchored oligo-dT primers were obtained from Oligos Etc Inc. (Wilsonville, Oregon). The single-strand DNA template was gel purified. The sequences for all the primers are listed in Table 2.

Generation of double-strand DNA template. To generate double-stranded DNA used as the templates for the synthesis of *in vitro* transcripts, the single-stranded DNA was amplified by PCR with 5′ T7M13 primer and 3′ T3 primer. The resulting double-stranded DNA templates contained a T7 promoter sequence at its 5′ end.

Preparation of *in vitro* transcripts. *In vitro* transcripts were prepared following the manufacturer's protocol (Promega, 1991, Madison, WI). In brief, 100 µl of reaction mixture contained 20 µl of 5x buffer (200 mM Tris-HCI, pH7.5, 30 mM MgCl2, 10 mM spermidine, 50 mM NaCl), 10mM DTT, 120 units of RNase inhibitor (Promega, Madison, WI), 1 mM NTP mixture, 50 units of T7 RNA polymerase (Promega, Madison, WI), 100 ng of double-strand DNA templates, and 32 µl of H_2O. The reaction was carried out at 37°C for 2 hours. Afterwards, 100 units of RNase-free DNase I (Pharmacia, Piscataway, NJ) was added to the reaction mixture at 37°C for another hour to digest DNA templates. The synthesized transcripts were extracted with phenol-chloroform, precipitated, washed, and dissolved in DEPC-treated H_2O.

The purity of *in vitro* transcripts was confirmed by direct PCR amplification of RNase treated *in vitro* transcripts with negative result.

cDNA synthesis and PCR amplification. cDNAs were synthesized with each anchored oligo-dT primer and MMLV reverse transcriptase (Promega, Madison, WI) or AMV reverse transcriptase (InvitroGen, Carlsbad, CA). Fifty μl of reaction mixture consisted of 5 μl of 5x buffer from each supplier, 2 mM dNTP mixture, 25 mM DTT, 20 units of RNase inhibitor, 50 ng of *in vitro* transcripts, 50 ng of each anchored oligo-dT primer, 25 units of MMLV reverse transcriptase or 5 units of AMV reverse transcriptase, and 9 μl of H_2O. Each reaction with MMLV reverse transcriptase was carried out at 37°C for one hour, with AMV reverse transcriptase at 42°C for one hour. The resulting cDNA was diluted 6 times with H_2O. One μl of cDNA was used for PCR amplification. Fifty microliters of PCR mixture contained 5 μl of 10x PCR buffer (500 mM KCl, 100 mM Tris pH9.0, 1% Triton X-100, Promega, Madison, WI), 2.5 mM MgCl2, 0.5 mM dNTP mixture, 5 units of Taq polymerase (Promega, Madison, WI), 100 ng of M13 primers, 100 ng of SP6 primers, and 1 μl cDNA templates. PCR was carried out at 94 °C/10 seconds, 60°C/10 seconds, and 72°C/10 seconds for 28 cycles.

Direct PCR amplification. Each anchored primer was used in combination with an M13 primer to amplify the original single-stranded DNA templates directly. As each anchored primer had only 11 dT residues, low stringency PCR conditions were used for the first 5 cycles at 94°C/10 seconds, 40°C/10 seconds, and 72°C/10 seconds. The reaction was then continued for 30 cycles at 94°C/10 seconds, 60°C/10 seconds, and 72°C/10 seconds.

TA cloning. The PCR products were cloned directly into pCR2.1 vector (InvitroGen, Carlsbad, CA), and were transformed into JM109 competent cells. Positive clones were selected by direct PCR amplification of colonies with M13R primer and T7 primer located in the vector.

DNA sequencing. The PCR amplified products were purified with Microspin S-300 HR columns (Pharmacia, Piscataway, NJ), and sequenced with dRhodamine sequencing kit (PE Applied Biosystems, Foster City, CA) and M13 R primer. The

sequences were collected by an ABI 377 automatic Sequencer and analyzed by ABI seq. analysis 3.0 software (PE Applied Biosystems, Foster City, CA).

Construction of cDNA library. mRNA was isolated from total RNA of HL60 cells with Dynal oligo-dT beads (Dynal, Oslo, Norway), following the manufacturer's protocol. The quality and purity of mRNA samples were checked on agarose gels. The cDNA library was generated with SaverTimer cDNA synthesis kit (Pharmacia, Piscataway, NJ) following the manufacturer's protocol, except two sets of mixed anchor primers were used for the reaction. Set 1 consisted of A-, G-, and C-anchored oligo-dT primers. Set 2 consisted of A-, G-, AC-, GC-, and CC-anchored primers. Double-stranded cDNA was cloned into pBS KS(-) vector. Positive clone identification and sequencing analysis were the same as above.

Example 2: Results

Experimental design. In order to generate precise information and to simplify the experimental process, an *in vitro* system was designed as shown in FIG. 2. The components in this system include DNA templates for generation of *in vitro* transcripts, 3′-anchored primers for cDNA synthesis, reverse transcriptases for reverse transcription, and primers for PCR amplification. The *in vitro* transcripts generated contained (i) 100 A residues to mimic polyA sequences in real mRNA templates; (ii) randomized nucleotides with A, G, or C at the first position immediately 5′ of the polyA sequences; (iii) randomized nucleotides with A, G, C, or T at the second position 5′ of the polyA sequences. These random nucleotides cover all the possible 2 nucleotide combinations at these two positions within any mRNA population (Table 2). The *in vitro* transcripts served as the templates for cDNA synthesis with each anchored oligo-dT primer and reverse transcriptase. The resulting cDNAs contained M13 sequence at their 5′ end and SP6 sequence at their 3′ end attached to the anchored oligo-dT primer. The single-stranded cDNAs were then converted to double-stranded cDNAs through PCR amplification with M13 primer and SP6 primer and were cloned for sequencing analysis. To exclude the possibility that Taq polymerase was responsible for the observed results, the single-stranded DNA templates were amplified directly by PCR with M13 primer, each anchored oligo-dT primer and Taq DNA polymerase.

Table 2

List of the Primers Used for the Analysis

Primers	Sequences
DNA template	5' GTAAA ACGAC GGCCA GTACG N* B** $(A)_{100}$ CTTTA GTGAG GGTTA ATTTC 3'(SEQ ID NO:1)
T7/M13 primer	5' CGTAA TACGA CTCAC TATAG GGGTA AAACG ACGGC CAGTA CG 3' (SEQ ID NO:2)
T3 primer	5' GAAAT TAACC CTCAC TAAAG 3'(SEQ ID NO:3)
T7 primer	5' CTAAT ACGAC TCACT ATAGG GC 3'(SEQ ID NO:4)
SP6 primer	5' CGATT TAGGT GACAC TATAG 3'(SEQ ID NO:5)
M13 R primer	5' CAGGA AACAG CTATG AC 3'(SEQ ID NO:6)
M13 primer	5' GTAAA ACGAC GGCCA GTACG 3'(SEQ ID NO:7)
One base-anchored oligo-dT	
SP6 T11A	5' CGATT TAGGT GACAC TATAG T_{11} A 3' (SEQ ID NO:8)
SP6 T11G	5' CGATT TAGGT GACAC TATAG T_{11} G 3' (SEQ ID NO:9)
SP6 T11C	5' CGATT TAGGT GACAC TATAG T_{11} C 3' (SEQ ID NO:10)
Two base-anchored oligo-dT	
SP6 T11CA	5' CGATT TAGGT GACAC TATAG T_{11} C A 3' (SEQ ID NO:11)
SP6 T11CG	5' CGATT TAGGT GACAC TATAG T_{11} C G 3' (SEQ ID NO:12)
SP6 T11CC	5' CGATT TAGGT GACAC TATAG T_{11} C C 3' (SEQ ID NO:13)
SP6 T_{11}CT	5' CGATT TAGGT GACAC TATAG T_{11} C T 3' (SEQ ID NO:14)

* N=A, or G, or C, or T

**B=A, or G, or C

There are only two possible outcomes for the resulting sequences. First, if the anchored oligo-dT primer annealed to the correct location within the transcript and is extended by reverse transcriptase, the resulting cDNA should contain only 11 dT sequences from the primer. However, if the anchored oligo-dT primer annealed randomly to any location along the polyA strand is extended by reverse transcriptase, the resulting cDNA will have between 12 and 100 dTs. Thus, the number of dT residues is an indication of the fidelity of the reverse transcriptases in the initiation of cDNA synthesis using anchored oligo-dT primers. In this reaction, the initiation of reverse transcription from the correctly annealed anchor primers contained only 11 dT

residues from the primer itself, whereas a long dT strand was generated from the unpaired anchor in the anchored primers along the polyA strand of the RNA. FIG. 3.

Sequence pattern from one-base-anchored primers. Three one-base-anchored oligo-dT primers T11A, T11G and T11C, and two widely used reverse transcriptases, MMLV reverse transcriptase and AMV reverse transcriptase, were used for reverse transcription. cDNAs generated were amplified, cloned and sequenced (Table 3). The results showed that 97% of clones generated with T11A and T11G primers by MMLV reverse transcriptase gave the correct cDNA sequences with only 11 dT residues. Surprisingly, from these two primers and AMV reverse transcriptase, only 53% and 47% clones showed correct sequences with 11 T residues. Even worse, with the T11C primer, 62% of the clones from MMLV reverse transcriptase and 73% of the sequences from AMV reverse transcriptase had various lengths of dT residues longer than 11 dT. With all these anchored primers, Taq DNA polymerase generated 100% correct sequences with only 11 dT residues.

Table 3
Sequences Resulting from One Base-Anchored Primers

3' primers	MMLV RT			AMV RT			Taq POL		
	Total Clones	Correct	Mis-extend	Total Clones	Correct	Mis-extend	Total Clones	Correct	Mis-extend
T11A	37	36 (97%)	1 (3%)	30	16 (53%)	14 (47%)	36	36 (100%)	0
T11G	33	32 (97%)	1 (3%)	34	16 (47%)	18 (53%)	31	31 (100%)	0
T11C	61	23 (38%)	38 (62%)	64	17 (27%)	47 (73%)	34	34 (100%)	0

Sequence pattern from one and two-base-anchored primers. To determine if the mis-extension of C-anchored oligo-dT primer could be corrected, two-base-anchored oligo-dT primers were designed with an additional nucleotide A, G, C, or T anchored to the C anchor of the dT11C primer, and used for reverse transcription. The results showed that dT11CA and dT11CG decreased the error rate to nearly 0% for both MMLV reverse transcriptase and AMV reverse transcriptase (Table 4). dT11CC decreased the error rate from 62% to 33% for MMLV reverse transcriptase, and from 73% to 19% for AMV reverse transcriptase. Taq polymerase generated 100% correct clones with these three primers. However, with dT11CT, an 84% error rate was observed for MMLV reverse transcriptase, and 97% for AMV reverse transcriptase. With Taq polymerase, the error rate was 67%.

Table 4
Sequences Resulting from Two Base-Anchored Primers

3' primers	MMLV RT			AMV RT			Taq POL		
	Total Clones	Correct	Mis-extend	Total Clones	Correct	Mis-extend	Total Clones	Correct	Mis-extend
T11CA	32	31 (97%)	1 (3%)	36	34 (94%)	2 (6%)	35	35 (100%)	0
T11CG	32	32 (100%)	0	36	36 (100%)	0	30	30 (100%)	0
T11CC	30	20 (67%)	10 (33%)	36	29 (81%)	7 (19%)	30	30 (100%)	0
T11CT	32	5 (16%)	27 (84%)	36	1 (3%)	35 (97%)	33	11 (33%)	22 (67%)

Sequence pattern from one and two-base-anchored primers with mRNA sample. To see if the same pattern can be generated with an actual mRNA population, anchored primers were mixed and used to generate cDNA libraries with MMLV reverse transcriptase. As shown in Table 5, the error rate was 56% with the mixture of dT11A-, dT11G-, and dT11C-anchored primers; with the mixture of dT11A-, dT11G-, dT11CA-, dT11CG-, and dT11CC-anchored primers, the error rate decreased to 14%.

Table 5

Sequences from mRNA, Mixed Anchored Primers and MMLV RT

Set 1 Mixture*		Set 2 Mixture**	
Correct	**Mis-extend**	**Correct**	**Mis-extend**
11 (44%)	14 (56%)	18 (86%)	3 (14%)

* Set 1 consisted of A-, G-, and C-anchored primers
**Set 2 consists of A-, G-, AC-, GC-, and CC-anchored primers

Example 3: Discussion

The presence of poly-dA/dT sequences in cDNA templates can be one of the major causes that effects the efficiency of gene identification. Results from a SAGE analysis detected 48,741 genes from normal and cancer cells, of which 41,882 genes (86%) were expressed at fewer than 5 copies per cell comprising only 25% of total mRNA. The remaining 6,859 genes (14%) were expressed from 6 to 5,300 copies per cell comprising 75% of the total mRNA (Zhang *et al.*, 1997). The hybridization with poly-dA/dT plus populations will certainly remove many lower abundant copies by the high abundant templates due to the random hybridization between poly-dA and poly-dT sequences (FIG. 4). This phenomenon should not happen in the poly-dA/dT minus cDNA populations. Moreover, generation of poly-dA/dT minus cDNA also is a pre-requirement for the identification of differentially expressed genes in a widely used technique, differential display (FIG. 5).

With one-base-anchored oligo-dT primers and MMLV reverse transcriptase for reverse transcription, A and G anchors generally ensure the correct initiation of

cDNA synthesis. C anchor leads to a high rate of non-specific initiation, resulting in variable lengths of poly-dT sequence in the generated cDNAs. These three one-base-anchored oligo-dT primers all lead to a high rate of longer poly-dT sequences in the generated cDNAs by AMV reverse transcriptase, with the highest error rate of 70% from the C-anchored primer. Direct PCR™ amplification of the DNA template with each anchored primers and Taq polymerase showed 100% correct sequences. This proves that the variable lengths of polydT sequences were generated from the reverse transcription, due to the non-specific initiation of cDNA synthesis from the unpaired anchors along the polyA strand by the reverse transcriptases. It has been observed that reverse transcriptases from retroviruses have mispair extension capacity in both DNA-dependent and RNA-dependent DNA syntheses which contributes to the high mutation rates of retrovirus replication (Abbotts *et al.*, 1991; Yu and Goodman, 1992; Bakhanashvili and Hizi, 1993). The inventors' observations indicate that a high rate of non-specific initiation exists in the RNA-dependent DNA synthesis by MMLV and AMV reverse transcriptases with anchored oligo-dT primers. The theoretical coverage of each one-base-anchored primer is 33.3% of the total mRNA population. The actual rate of incorrect initiation through C-anchored primer would be higher than that from A- and G-anchored primers, because more C-anchored primers which annealed randomly on polyA sequences will be extended by reverse transcriptase, comparing only one A- and G-anchored primer being extended. This may explain why 45% of the identified clones were from the C-anchored primer rather than the expected 33% in the cDNA library generated with the mixture of one-base-anchored primers, assuming the A, G and C are equally distributed in the last position before the polyA sequences among the expressed sequences.

Adding a second nucleotide to the C-anchored oligo-dT primers resulted in significant changes in the pattern. CA- and CG-anchored primers prevent the non-specific initiation from the C-anchored primer for both MMLV reverse transcriptase and AMV reverse transcriptase; using a CC-anchored primer also significantly reduced the error rate for both reverse transcriptases, even through a certain number of non-specific products still exist. With a CT-anchored primer, however, the error rate reached 80% and 97% for MMLV reverse transcriptase and AMV reverse transcriptase, respectively. More importantly, the CT-anchored primer generated a

67% error rate in DNA synthesis by Taq DNA polymerase, indicating that CT-anchors lead to the non-specific initiation in both RNA-dependent and DNA-dependent DNA synthesis. This additional T residue can anneal with an A residue in the mRNA polyA strand, much like dT's in the oligo-dT primer annealed along the polyA strand. The inclusion of a third nucleotide to the CT anchors (*i.e.*, CTA-, CTG-, and -CTC) will not significantly improve this situation. These three-anchored primers will essentially function as the single base-anchored primers because the CT-anchored primer functions similar to oligo-dT primer. CTT-anchored oligo-dT primer will function as a CT-anchored oligo-dT primer, and it too can not be used. The exclusion of a CT-anchored primer means that mRNA ended with AG before the polyA sequence will not be included, with a theoretically coverage of 1/12 (8.3%) of the total expressed sequences. This would appear to be a small price to pay for the substantial recovery of sequence information from genes expressed at a lower level. MMLV reverse transcriptase should be used rather than AMV reverse transcriptase, as it correctly initiates cDNA synthesis with A and G anchored oligo-dT primers.

Thus, from the data presented here, the inventors conclude that the minimal optimal number of anchored oligo-dT primers for generation of a poly-dA/dT minus cDNA population should be A-, G-, CA-, CG-, and CC-anchored oligo-dT primers, together with MMLV reverse transcriptase. This combination will provide the specificity, the simplicity, and the maximal coverage of the expressed sequences. With this strategy, the inventors routinely obtain cDNA populations which are about 85% poly-dA/dT minus. Removal of a CC-anchored oligo-dT from the combination may further increase the rate of poly-dA/dT minus cDNAs, but will decrease the coverage of expressed sequences to 83.4%.

Due to its simplicity and potential for covering most of the expressed genes, the differential display technique has been widely used to identify differentially expressed genes (Liang and Pardee, 1992). However, a major problem for this technique is the high rate of false positives (Sun *et al.*, 1994). Based on a study that a C-anchored primer specifically amplified DNA template in PCR™ by Taq polymerase, it was concluded that all three one-base-anchored primers provide specificity for reverse transcription, resulting in the three poly-dA/dT minus cDNA subpopulations (Liang *et al.*, 1994). From the inventors' observations, they believe

that this conclusion is incorrect and that it has serious implications for the use of the present differential display strategy. The inventors conclude that the high false positive rate in differential display is largely due to the inclusion of poly-dA/dT sequences that originated from reverse transcription with the C-anchored oligo-dT primer (Liang *et al.*, 1994). When the cDNA products generated from C-anchored oligo-dT primers are amplified and displayed on gel, many fragments will appear as "differentially expressed" genes distributed at various unique locations because of their different sizes. However, many of them are in fact false positives of which the differences in size are due to the inclusion of the different length of poly-dT sequences. Because the error is generated at the very beginning, efforts of many investigations focusing on downstream modifications to correct the errors are destined to be unsuccessful. A similar situation arises from TC-anchored primers when two-base-anchored oligo-dT primers are used (Liang and Pardee, 1992). The inventors conclude that the principles they have established for the generation of poly-dA/dT minus cDNA population should also be applied to differential display. The minimal number of anchored oligo-dT primers for differential display technique should be five, including A-, G-, CA-, CG-, and CC-anchored oligo-dT primers, which result in a theoretical coverage of 91.7% of the total expressed genes. The C-anchored and TC-anchored oligo-dT primers should not be used MMLV RT instead of AMV RT should be used for reverse transcription. The inventors' conclusion should be immediately applicable to correct the problem inherent in the current version of the differential display technique.

Example 4: Materials and Methods

1. Cell Culture

HL60 cells were cultured at 37°C in RPMI 1640 medium with 10% fetal calf serum. Cells were harvested at exponential phase for RNA isolation.

2. mRNA Isolation

Total RNA was isolated from HL60 cells with Trizol reagent (Life Technologies, Gaithersburg, MD) following the manufacture's instruction. The isolated RNA was then treated with DNase I, and run on agarose gels to evaluate the quality of the RNA. mRNA was isolated from total RNA with Dynal dT_{25} (Dynal,

Oslo, Norway) following the manufacturer's protocol. mRNA samples were run on an agarose gel to determine purity, quantified at OD260, and stored at -70°C.

3. cDNA Synthesis

Double-stranded cDNAs were synthesized with a cDNA synthesis kit (Life Technologies) following the manufacture's protocols, except that a mixture of three anchored and biotinylated primers for the reverse transcription reaction was used. The sequences of these three primers are 5′ biotin-TTTGCATGCTCGAG-T_{16}-A/G/C **(SEQ ID NO:15 - SEQ ID NO:17)**. The synthesized double-stranded cDNAs were treated with phenol-chloroform, precipitated, washed and dissolved in TE buffer.

4. 3′ cDNA recovery

Double-stranded cDNAs were digested with *Nla*III at 37°C for 2 h. 3′ cDNA was recovered with Dynal M280 avidin beads (Dynal, Oslo, Norway) according to the manufacturer's protocol. After washing away the unbound fragments, the bound 3′ cDNA was released from the beads by mixing with phenol at 65°C for 30 min and vortexing at full speed for 10 min. Recovered 3′ cDNAs were precipitated, washed, and dissolved in TE buffer. The purified 3′ cDNA was further digested with *Sph*I in order to generate a CATG end within the RT primer sequence for adapter ligation.

5. Ligation of Adapters

The 3′ cDNAs were divided into two groups. One was designated tester, and the other driver. The tester was divided further into two sets for ligation to adapter A or adapter B. The sequence of adapter A was, sense: 5′ ATA CGA CTC ACT ATA GGG CTC GAG CGG CCG CAT ATG GGA CAT G 3′ **(SEQ ID NO:18)**; antisense: 5′ TCC CAT ATG C 3' **(SEQ ID NO:19)**. The sequence of adapter B was, sense: 5′ ATA CGA CTC ACT ATA GGG CAG CTC GCC GGC GTA TAG GGA CAT G 3′ **(SEQ ID NO:20)**, antisense: 5′ TCC CTA TAC G 3′ **(SEQ ID NO:21)**. The primers were modified from the original primer sequences (Siebert *et al.*, 1995) in such a way that its 5′ part was T7 promoter sequences and its 3′ part was an *Bsm*FI/*Nla*III recognition sequence. *Bsm*FI site can be used to obtain a tag sequence from templates for SAGE analysis (Velculescu *et al.*, 1995). The ligation reactions were carried out at 16°C overnight. *Nla*III digested pBR322 DNA was used as a control. The ligation

efficiencies were monitored by PCR™ with T7 primer 5′ CTA ATA CGA CTC ACT ATA GGG C 3′ **(SEQ ID NO:31)**.

6. 3′ cDNA subtraction

The subtraction reaction was performed following the protocol for suppression subtraction hybridization (Diatchenko *et al.*, 1996). Twenty ng of cDNA was used as the tester. Different ratios between tester and driver were set from 1:0 to 1:35 for both tester A/driver and tester B/driver. After denaturing at 98°C for 2 minutes, the first hybridization was carried out for 10 h at 68°C. After mixing sample A and sample B, the second hybridization was performed at 68°C for another 10 h. Samples were then diluted for suppression PCR™ amplification.

7. Suppression PCR™

Suppression PCR™ was performed using T7 primer. The reactions were first incubated at 74°C for 5 min to extend the 3′ end of the adapter in order to generate the templates for T7 primer binding in the PCR™ reaction. PCR™ was performed with 94°C/10 sec, 66°C/20 sec and 72°C/20 secs. pBR322 DNA with adapter A, adapter B and adapter A/B were set as the control to monitor the suppression effects of the reaction. After every 2 cycles starting from the 22nd cycle, the patterns of amplification were checked by loading 5 µl of the PCR™ samples on an agarose gel. The amplifications were stopped when clear signals were seen on the gel (A/B), but the noise signals represented on the control reactions (A or B) were not significantly amplified. The PCR™ products then were purified and adjusted to the same concentration.

8. Multiplex quantitative PCR™

β-actin, *HSC70* and *HSP75* were selected as the indicators for the determination of subtraction efficiency. β-actin is expressed at abundant level and is used widely as a control in the analysis of gene expression. *HSC70* and *HSP75* were selected as the intermediately expressed genes. Control templates, homologous to but shorter than the wild-type templates, were generated through amplifying the corresponding cDNA with the 3′ primer and a truncated 5′ primer (Table 6). These primers are located downstream of the last *NIa*III site of these cDNAs. The 5′

truncated primer contained the same 5′ primer sequences as the wild-type template but was ligated further 3′ sequences resulting in a gap. The gap is 20 base pairs, 20 base pairs and 10 base pairs for β-actin, *HSP75* and *HSC70,* respectively. The templates generated by these primers have the same sequences as wild-type but are shorter because of the gap formed by the 5′ primer. The multiplex PCR™ reactions were performed by adding the samples to the reaction mixtures containing the 5′ and 3′ primers from all these genes and defined amounts of control templates, and α^{32}P-dCTP. The PCR™ conditions were 94°C/10 sec, 55°C/20 sec and 72°C/20 sec for 38 cycles. The PCR™ products were fractionated on a 5% denaturing gel and exposed on a PhorsphaImage plate. The signal intensities were measured by ImageQuant (Molecular Dynamics, CA).

Table 6

Primers Used for Quantitative PCR

	Primers Sequences
β-Actin	
Sense 1 **SEQ ID NO:22**	TGTTACAGGAAGTCCCTTGCTTCTCTCTAAGGAGAATGGC
Sense 2 **SEQ ID NO:23**	TGTTACAGGAAGTCCCTTGC
Antisense **SEQ ID NO:24**	TAAGGTGTGCACTTTTATTC
HSC70	
Sense 1 **SEQ ID NO:25**	CCAGGAGGAATGCCTGGGGTGGTGGAGCTCCTCCT
Sense 2 **SEQ ID NO:26**	CCAGGAGGAATGCCTGGG
Antisense **SEQ ID NO:27**	TTAATCAACCTCTTCAATGG
HSP75	
Sense 1 **SEQ ID NO:28**	AGATAAAGGCACAAGACGTGTCTTCTGGTGGATTAAGCAA
Sense 2 **SEQ ID NO:29**	AGATAAAGGCACAAGACGTG
Antisense **SEQ ID NO:30**	GCAGGTAATTGGTCCTTGAA

The ratio between wild-type and control templates was determined for each cDNA. The subtraction efficiencies were determined by comparison of these ratios among different samples.

9. cDNA Sequencing and Sequence Alignments

The amplified cDNAs from suppression PCR™ were directly cloned into TA vector (InvitroGen, Carlsbad, CA). Sequencing reactions were performed using Ml3 reverse primer and ABI cycle sequencing kit, and sequences were collected on an ABI377 autosequencer (Applied BioSystems, Foster City, CA). For database alignment, each sequence was first matched to GenBank databases by BLAST search; if no match was found, the same sequence was used to match dbEST database. If no match was found in either database, the sequence was designated a novel sequence.

Example 5: Results

The strategy used is illustrated in FIG. 1. To validate the system, the same mRNA was used for the preparation of both tester and driver, and no driver was used to subtract the control sample. The final results should indicate whether the system functions as predicted, *i.e.*, the abundant and intermediate abundant sequences should decrease, and the proportion of rare abundant sequences should increase.

To generate a cDNA source for isolation of 3′ fragments from all the expressed genes, the double-stranded cDNAs were digested by the restriction enzyme *Nla*III. *Nla*III recognizes the CATG sequence that occurs on average every 256 base pairs. To verify the actual size distribution after *Nla*III digestion, the digested cDNA fragments were run on an agarose gel. As shown in FIG. 6, the fragments were primarily centered between 300 to 500 bp.

By using avidin beads, the 3′ biotinylated cDNAs were isolated from the total *Nla*III digested cDNA. The recovered 3′ cDNA all included a CATG site at their 5′ end generated by *Nla*III digestion. To generate a CATG site at their 3′ end for the addition of the adapter, an *Sph*I site (GCATGC) was designed in the RT primer. Even though it contained CATG that is an *Nla*III cleavage site, this site cannot be digested

by *NIa*III due to shortness of 3' end flanking bases. The recovered 3' cDNAs were further digested by *Sph*I.

To determine the most efficient subtraction protocol, a series of ratios between tester and driver were set. The first round subtraction was used to eliminate the abundant and intermediate abundant templates, followed by the second round hybridization for annealing the unsubtracted templates. The unsubtracted cDNA was selectively amplified by suppression PCR™.

To determine the subtraction efficiency, the level of redundancy of the test genes was compared between the control and subtracted samples using a multiplex quantitative PCR™ assay (FIG. 7 and Table 7). β-actin was selected as a marker for abundant copies. *HSC70* and *HSP75* were selected as the representatives of intermediate abundant copies being expressed at hundred copies per cells in human cell lines. The β-actin level decreased 76-fold at a ratio of 1:35; *HSC70* and *HSP75* copies were hardly detectable after a ratio of 1:15.

Table 7

Subtraction Efficiency for β-Actin Templates

tester:driver	relative:fold
1:0	1
1:15	4
1:25	23
1:35	76

A total of 93 clones were sequenced, including 41 clones from control sample and 52 clones from subtracted sample at ratio of tester:driver of 1: 35. (Table 8 and Table 9).

Table 8

Sequence Alignment of Control 3' cDNA Clones

Clones	GenBank+EMBL+DDBJ	EST Database	UniGene	Plasmid ID
1	ribosomal L5			
2	ribosomal S28			
3	elongation factor 1-alpha			
4	--	AA374089	--	178491
5	mitochondrial tRNAs and partial proteins 4 & 5			
6	ribosomal S3			
7	ribosomal protein L23a			
8	BN51			
9	KIAA0002			
10	DNA polymerase delta small subunit			
11	prothymosin alpha			
12	elongation factor 1-alpha			
13	elongation factor 1-alpha			
14	--	D45527	bleomycin hydrolase	lg1240
15	--	AA526048	similar to hypotheti	982571

Clones	GenBank+EMBL+DDBJ	EST Database	UniGene	Plasmid ID
			cal 75.2 kD protein	
16	cytochrome c oxidase SII			
17	ribosomal protein S28			
18	leucine-rich protein			
19	leucine-rich protein			
20	--	F20343	ribosomal protein S12	036-X3-02
21	T-cell cyclophilin			
22	elongation factor 1-alpha			
23	elongation factor 1-alpha			
24	--	T23659	--	b4HB3MA
25	elongation factor 1-alpha			
26	ribosomal protein S8			
27	ubiquitin activating enzyme E1			
28	--	W07352	elongation factor 1-alpha	300651
29	cytochrome c-1			
30	ribosomal protein S28			
31	ribosomal protein S24			

Clones	GenBank+EMBL+DDBJ	EST Database	UniGene	Plasmid ID
32	--	--		
33	ribosomal protein L37a			
34	--	--		
35	mitochondrial genome X62996			
36	--	N75815	ribosomal protein S23	300310
37	ribosomal protein L37a			
38	vacuolar H(+)-ATPase subunit			
39	elongation factor 1-alpha			
40	23 kD highly basic protein			
41	23 kD highly basic protein			

Table 9

Sequence Alignment of Subtracted 3′ cDNA Clones (Tester:Driver=1:35)

Clones	GenBank+EMBL+DDBJ	EST database	UniGene	Plasmid ID
1	--	N75293	Human Cbf5p homolog	298739
2	ribosomal protein L17			
3	--	T23947	--	HB3MA-4(3')

Clones	GenBank+EMBL+DDBJ	EST database	UniGene	Plasmid ID
4	--	AA142952	Weakly similar to EGF-R SUBSTRATE 15	504699
5	--	--	--	
6	--	AA256523	--	682558
7	--	AA570755	Weakly similar to MDR-1	914430
8	--	AA427866	--	771016
9	--	AA218963	--	650193
10	--	W69264	--	343643
11	--	--		
12	Human Hlark mRNA			
13	mitochondrial ubiquinone-binding protein			
14	eosinophil granule major basic protein			
15	--			
16	LLRep3			
17	TI-227H			
18	--	H07593	--	--
19	ribosomal protein L37a			
20	LLRep3			

Clones	GenBank+EMBL+DDBJ	EST database	UniGene	Plasmid ID
21	--	AA159050	Glucose phosphate isomerase	591011
22	--	AA310374	--	180113
23	--	--		
24	SnRNP core protein Sm D3			
25	transferrin receptor			
26	--	R48155	--	153719
27	--	AA524203	--	936933
28	KIAA0174 gene			
29	--	C18312	mitochondrial NADH Fe-S protein 8	18312
30	acidic ribosomal phosphoprotein P2 mRNA			
31	actin-related protein Arp3			
32	--	--		
33	--	T24112		Cot274
34	ribosomal protein L5			
35		AA583472	--	1086940
36		AA583472	--	1086940
37	human protective protein			
38	--	AA306845	Highly similar to	160936

Clones	GenBank+EMBL+DDBJ	EST database	UniGene	Plasmid ID
			AUP46 precursor	
39	ribosomal protein S17			
40	ribosomal protein S28			
41	--	--		
42	mitochondrial ubiquinone-binding protein			
43	--	H05063	--	43295
44	ferritin light subunit mRNA			
45	elongation factor 1-alpha			
46	NADH-ubiquinone			
47	--	AA484025	Ribosomal protein L19	910266
48	--	--		
49	--	AA156023	Ribosomal protein S8	590124
50	--	--		
51	ribosomal protein L27a mRNA			
52	--	--		

Example 6: Discussion

Analysis of gene expression has moved rapidly from classical studies on a single or a few genes toward genome-wide studies on multiple genes. As most of the traditional techniques have very limited capacity for analysis on such a scale, and

current techniques for genome-wide studies of gene expression have various limitations, the inventors developed this new procedure by integrating different parts of several techniques into a linear system. The data presented here show that the inventors substantially achieved their objectives.

The unique features of this system include the following. 1) Use anchored oligo dT primer to generate polydA dT minus cDNA to prevent random poly-dA/poly-dT hybridization between the templates in the subtraction process. This feature will largely conserve the rare copies after subtraction for gene identification. 2) The templates are located at the 3′ end and contain mostly 300 to 500 base pairs. This feature guarantees the maximal use of EST information for gene identification. 3) The redundancy can be largely decreased through the subtraction reaction. By using quantitative multiplex PCR™ to verify the subtraction efficiency, the mRNA requirement can be significantly decreased and the subtraction efficiency can be precisely determined. The inclusion of the SAGE technique further significantly decreases the number of sequencing reactions required for genome-wide scanning.

The inventors' data show that these features provide a cDNA population in which (i) much of the sequence redundancy can be significantly reduced and the rare templates can be enriched, (ii) the sequences identified can be used directly to match the EST database for gene identification, and (iii) any sequences unmatched to the EST database are likely to be *bona fide* novel sequences not yet existing in the database. In the report of Diatchenko *et al.* (1996), 55 out of 62 sequences identified through subtraction suppressive PCR™ (SSH) were considered as "novel" sequences because they were not matched to databases. However, as these sequences could have came from anywhere in the cDNA template, these supposedly new sequences could represent genes whose 3′ or 5′ sequences were already in the EST databases. To analyze these "novel" genes further, one would need to use traditional methods of screening cDNA libraries in order to clone these allegedly "novel" cDNAs. The inventors' approach has a very high likelihood of distinguishing the identified sequences as existing EST sequences or real novel sequences.

Compared to the 5′ EST sequences, the 3′ EST sequences are the most reliable ones as the cDNAs generated in reverse transcription are frequently unable to reach

the 5′ end of the templates due to the existence of mRNA secondary structure. Several approaches have been developed in recent years for using the 3′ portion of cDNA for gene identification (Velculescu *et al.*, 1995; Ivanova *et al.*, 1995; Kato, 1995; Prasher *et al.*, 1996). The advantage of focusing on 3′ sequences is well demonstrated by the inventors' system. First, it provides a better representation of genes, as the 3′ part is the highly heterogeneous portion of the gene. Second, because sequences are relatively short for each gene, the probability of false hybridization among different genes will be decreased which further increases the specificity of the subtraction reaction. Third, each expressed gene in the analysis has only one marker represented by its 3′ sequence, thus avoiding the uncertainty that multiple sequences may be generated for the same gene due to analysis of the 5′ and 3′ portions when regular cDNA libraries are used. Fourth, and most importantly, it guarantees that the sequences identified will have the highest likelihood of matching to any existing EST sequences. The size of 300-500 base pairs parallels the length of EST sequences, and is sufficient as a specific marker for each gene. For these novel sequences, techniques such as 5′ RACE can be used to obtain more 5′ sequences if necessary (Bertling et al., 1993).

A potential application of the IPGI technique would be for the EST project. The number of EST sequences currently is increasing rapidly, but the number of unique genes identified from these sequencing efforts is diminishing (http://www.ncbi.nlm.nih.gov/ ncicgap/gene-discovery.html). This could indicate that most of the expressed genes might have been identified through the current EST project. However, when comparing data from SAGE analysis and the EST project, it is interesting to note that nearly half of the sequences identified by SAGE have no match in databases including EST database (Zhang et al., 1987), indicating the potential of more sequences waiting to be identified. Many of these unidentified sequences may be expressed at a very low level. The libraries used for the EST project are generated by oligo-dT priming in reverse transcription which generates cDNAs all containing poly-dA/poly-dT sequences at their 3′ ends, followed by normalization/subtraction before being used for sequencing reaction (Bonaldo et al., 1996). The random hybridization between poly-dA and poly-dT sequences in the normalization/subtraction process may lead to heavy loss of the rare copies by the

abundant copies. This can be one of the major reasons why the current EST project has difficulty in identifying more genes, particularly genes expressed at rare level. With the approaches described here, it should be possible to generate the libraries with a better representation of the rare transcripts. This may significantly increase the rate of novel sequence identification.

The IPGI procedures also may be applied to CGAP project. The priority in the current CGAP project is to index all genes expressed in primary tumors (Strausberg et al., 1997). Due to the large size of the human genome and the redundancy of the expressed transcripts, it is difficult, if not impossible, to identify all the expressed genes by direct sequencing of the primary cDNA library from each tumor. The normalization/subtraction strategy would be a necessary step in order to decrease the redundancy for the analysis. On the other hand, it is very likely that in many tumor cells, the abnormally expressed genes account for only a small portion of the total expressed genes, and the majority of expressed genes would be the same as these expressed in normal cells (Zhang et al., 1997). The EST project provides a large number of sequences expressed in normal cells. Maximal use of EST information will significantly decrease the cost for indexing genes expressed in tumor cells. The features of IPGI described earlier make it an ideal choice for the CGAP project: (i) sequences generated through the IPGI technique provide a high degree of completeness in covering most of the expressed templates, particularly the rare copies; (ii) only one unique 3′ marker for each gene will be generated, which increases the specificity for gene identification, and cuts the cost in half if regular libraries are used for 5′ and 3′ sequencing; (iii) the overall work can be significantly decreased through the normalization process; and (iv) the EST information can be maximally used.

In summary, the development of IPGI procedures provides a tool for genome-wide gene analysis. It should find wide application in functional genomic studies. Of equal importance is the fact that it can be used in standard molecular biology laboratories to address genome-wide questions heretofore unanswerable.

Example 7: Screening Poly dA/dT (-) Minus cDNAs for Gene Identification (SPGI)

The normalization/subtraction methods used to reduce the high-abundant copies involve the generation of double-strand hybrids containing high-abundant genes and their removal by hydroxyapatite absorption. However, during the normalization/subtraction process, random hybridization occurs between the poly dA and poly dT sequences in the 3' end of cDNA templates included by oligo dT priming. This results in the formation of tangled polydA/polydT double-strand hybrids independent of the sequence specificity (FIG. 4). Because double-stranded hybrids are removed, copies of many genes inappropriately annealed to the hybrids could be lost. Those lost copies will not be identified despite extensive sequencing efforts. This will affect particularly the identification of genes expressed at low levels. The inventors describe herein a method called screening poly dA/dT (-) minus cDNAs for gene identification (abbreviated to SPGI) to address these problems.

To test the formation of tangled poly dA/dT molecules, the inventors designed an *in vitro* model in which a single-strand synthetic DNA template containing 100 dA residues was subtracted with a cDNA pool containing long poly dT sequences, absorbed using hydroxyapatite, and subsequently quantified by quantitative PCR. The results show that the template was lost after these procedures, indicating that the formation of poly dA/poly dT hybrids during subtraction can indeed result in the loss of templates.

The inventors further reasoned that, if cDNA templates do not contain a long poly dA/dT sequence, those templates could be preserved after the subtraction. Such cDNA templates can be generated by using 3'-anchored oligo dT primers instead of regular oligo dT primers for reverse transcription. The assumption is that only the primers annealed to the 5' end of the mRNA poly A tail and its anchor nucleotide paired to the nucleotide immediately 5' of the poly A sequence could result in extension by reverse transcriptase. Primers annealed to other positions in the poly A sequence should not be extended, because the unpaired anchor blocks extension. These two features should prevent the cDNA from inclusion of a long poly dT sequence. In their initial attempts, the inventors frequently observed that many clones

still contained long poly dA/dT sequences despite the use of anchored oligo dT primers for reverse transcription.

The inventors examined systematically the pattern of cDNA synthesis with anchored oligo dT primers and reverse transcriptases. An *in vitro* transcript was synthesized to mimic mRNA templates, which contained 100 adenosine residues, randomized nucleotides of A, G, or C at the first position 5' of the poly A sequences, and randomized nucleotides of A, G, C, or T at the second position 5' of the poly A sequences. The distribution of these random nucleotides reflects all of the possible combinations at these two positions within natural mRNA populations. Thus, one base anchored and two base anchored oligo dT primers were used for the priming. There are two possible outcomes for a given cDNA clone after reverse transcription: either the clone contains 11 dA/dTs at its 3' end, derived from the anchored oligo dT primer annealed to the 5' end of the poly A sequences, or it contains a longer poly dA/dT sequence at its 3' end, resulting from the primer annealed randomly along the poly A sequences. The inventors classified the former as poly dA/dT(-) clone, and the latter as a poly dA/dT (+) clone. To their surprise, the results showed that the lengths of the poly dA/dT sequences in the cDNA clones were anchor nucleotide-dependent, and reverse transcriptase-dependent (FIG. 3, Tables 3, 4). The cDNA generated from dA- and; dG- anchored primers by MMLV reverse transcriptase were almost all poly dA/dT(-) clones. With AMV reverse transcriptase, however, nearly half of the clones were poly dA/dT (+). Most clones generated with a dC-anchored primer using either MMLV or AMV reverse transcriptase were poly dA/dT (+). Thus, the dC-anchored primer does not provide a discriminatory function for the synthesis of a poly dA/dT(-) clone. This feature contributes directly to the inherent problem of high false-positive rates of gene identification in the differential display technique. Because of the inclusion of a random length of poly dA/dT sequences at the 3' end of cDNA templates through the dC-anchored primer, the size of a particular cDNA template varies, which makes gene identification through gel fractionation highly unreliable. The addition of a second anchor to the dC-anchor alters this pattern. dCdA and dCdG anchors resulted in poly dA/dT(-) clones from both MMLV and AMV reverse transcriptases. The dCdC anchor moderately decreased the poly dA/dT (+) rate with both MMLV and AMV reverse transcriptases. With dCdT anchors, however, most

clones were still poly dA/dT (+). This primer also created 2/3 of poly dA/dT (+) clones in the control PCR with Taq polymerase, indicating the non-selectivity of this primer. The inclusion of additional nucleotides to the dCdT anchors does not improve this situation due to the non-specificity of dCdT.

The inventors conclude from this study that the application of dA-, dG-, dCdA-, dCdG-, dCdC-anchored oligo dT primers and MMLV reverse transcriptase provides the optimal combination for the generation of poly dA/dT(-) cDNAs, with the simplicity, specificity and maximal coverage of the expressed mRNAs. The recognition of the total expressed sequences with these primers is 91.7%, assuming the random distribution of A, G, C, and T in the last and second-last positions before the poly A sequences. The mRNAs ending with the nucleotides AG (which constitute about 8.3% of the mRNA) will not be included because of the exclusion of the dCdT-anchored primer. Thus, by applying this strategy, the inventors routinely obtain over 90% poly dA/dT(-) cDNA clones after reverse transcription with different mRNA samples.

The inventors next performed an experiment similar to the experiment designed to test whether the poly dA/dT(-) templates would be preserved. In this experiment, the driver cDNA used was a poly dT (-) population generated by the optimal combination of piers and reverse transcriptase, and the tester template contained only 16 dA. The results showed that the templates were largely retained after the treatments. This indicates that the exclusion of long poly dA/dT in the cDNA can indeed preserve templates upon subtraction.

To determine whether their *in vitro* data would predict greater efficient in retention of low abundant templates *in vivo*, the inventors compared their strategy with the standard approaches used in EST/CGAP projects. An mRNA sample from normal colonic epithelium cells was chosen for this companion, because the expression pattern of this sample had been analyzed extensively by the SAGE technology. Of 14,721 genes identified from 62,168 SAGE tags, over 70% were expressed at 5 copies or less per cell. The relative levels of a set of sequences expressed at low level in this sample were compared after subtraction and hydroxapatite absorption (Table 10). The results showed that the levels of these

sequences in the poly dA/dT(-) reactions were between 1.4- and 7.8-fold higher in 4 out of the 5 genes than that in the poly dA/dT (+) samples. The addition of a large excess of oligo dT_{20}, a method used routinely in normalization and subtraction in EST/CGAP projects in attempt to block the poly dA/poly dT hybridization, does not adequately preserve the low abundant copies, as shown as only a minor increase in two of the samples. No changes in the level of the AA297150 sequence in all three reactions is explained by the lack of a long poly A sequence in its original mRNA template. Table 10 shows the comparison of relative levels of the templates. The numbers within the parentheses are the original ratio between wild type and control amplicons. The numbers from the reaction of the poly dA (+) tester and poly dA/dT (+) driver (line 1) was set at 1.0. Numbers from other samples were normalized to this value for comparison.

Table 10

ITEM	GENE				
	A1193160	**AA435717**	**X03747**	**AA448394**	**AA297150**
poly dT (+)	1.0 (0.8)	1.0 (0.8)	1.0 (1.2)	1.0 (0.5)	1.0 (1.2)
poly dT (-)	4.9 (3.9)	7.8 (6.2)	2.6 (3.1)	1.4 (0.7)	1.0 (1.2)
poly dT (+)/ oligo dT_{20}	3.5 (2.8)	2.1 (1.7)	0.7 (0.8)	1.1 (0.6)	1.0 (1.2)

The inventors further verified their method by screening directly a normalized poly dA/dT(-) colon cDNA library for gene identification. Clones in the original library were also sequenced as a control to show the efficiency of this method. As shown in FIGs. 8A-8D, the rate of novel sequences identified in the normalized sample increased to 16% in a total of 193 clones analyzed, compared with 3% in the control sample of 109 clones. As an additional verification, SAGE tags were also collected from both samples. The alignment of SAGE tags with SAGE tag database shows that the number of novel SAGE tags in the normalized sample was much higher than that in the control sample (43% versus 16%). These data clearly indicate

that screening normalized/subtracted poly dA/dT(-) cDNA sample can generate a much higher degree of novel gene identification than is achieved with existing current approaches.

Thus, the inventors demonstrate that the presence of poly dA/dT sequences in cDNA templates leads to the loss of cDNA templates upon normalization and subtraction. It is very likely that this loss contributes in large measure to the low efficiency of novel gene identification in the current EST/CGAP projects. This obstacle can be overcome through applying SPGI technique. The rate of gene identification in the current CGAP is about 4.6%, based on generation of 1,000 EST sequences per day (http://www.ncbi.nlm.nih.gov//ncicgap/). If one assume that there are about 30,000 unknown genes, and all of which would eventually be identified through the current EST/CGAP approaches then about 652,174 sequences will need to be identified in about 652 days. However, if the rate can be increased to 16% with SPGI strategy, the total sequencing effort can be decreased to 187,500 which could be completed in 187 days. This would be a significant increase in the efficiency of novel gene identification. In addition, the SPGI technique is also applicable in the functional genomic studies with various higher eukaryotic systems in the post-genome era.

Materials and Methods of the SPGI Technique:

The steps involved in the SPGI method are depicted in the schematic in FIG. 9 and include the following steps:

1. Isolation of mRNA
2. Synthesis of poly dA/dT minus double-strand cDNA from mRNA
3. NIaIII digestion of double strand cDNA
4. Recovery of 3' cDNA pool
5. Cloning of the recovered 3' cDNA
6. Preparation of drivers for normalization reaction

7. Conversion of double strand to single strand cycle DNA

8. Performance of subtraction/ normalization and removal of the ds hybrids

9. Conversion of single strand to double strand

10. Preparation of plasmid

11. Sequence collection

12. Gene identification

The detailed protocol for SPGI describing the individual steps outlined above are as follows:

1. Isolation of mRNA involves

a. Isolation of total RNA was performed with Trizol solution (Life Technology), following manufacture's protocol, and quantified at O D 260.

b. Isolation of mRNA with Dynal dT25 beads by removal of 1 ml Dynal dT beads (5 mg) into an eppendorf tube and place in a microcentrifuge. The supernatant was removed re-suspend in 1000 ul 2 x binding buffer (20 mM Tris pH 7.5, 1 M LiCI, 2 mM EDTA), and placed in microcentrifuge. The supernatant was removed again and the beads were resuspend in 200 ul 2 x binding buffer. 200 ug/200 ul total RNA was added and mixed well. The mixture was incubated at room temp for 5 minutes and placed in a microcentrifuge for 1 minute. The supernatant was removed, washed 3x with 1,000 ul washing buffer (0.5 ml 1M tris pH8.0, 0.1 ml 0.5 M EDTA, 5 ml 1.5 M LiCI, H_20 till 50 ml) and the beads were resuspend with 20 ul elusion buffer (10 mM Tris pH 7.5). This was incubated at 65°C for 2 minutes and the supernatant was recovered by centrifugation. The beads were again resuspend in 10 ul of elusion buffer and keep at 65°C for 2 minutes and the supernatant was recover again. All the mRNA elutes were combined, checked on a gel, quantified at O D 260 and stored at -70°C. The beads were recondition with 500 ul 2 x binding buffer and are used for up to three times.

2. Synthesis of poly dA/dT minus double-strand cDNA

a. Rationale for RT primer designing:

-Attach biotin at the end of RT oligo to enable recovery of 3' cDNA with Dynal M280

-Incorporate a Not I site for removing Dynal 280 beads after recovering 3' cDNA and create a CCGG site for cloning into vector

-Use A/G, C-A/G/C anchored oligo dT primers for cDNA synthesis to avoid the inclusion of longer poly dA tail in the 3' cDNA pool (average cutting site of NIaIII is 256 base pairs, while average poly A tail is 50-250 bps).

b. Synthesize RT primers:

5' biotin-ATC TAG AGC GGC CGC-T16-R
5' biotin-ATC TAG AGC GGC CGC-T16-C-V

where R=A/G, V= *AIGIC* for random synthesis at this position

c. Prepare the mixture of anchored primers

A and G anchored primer to 2 ug/ul

C-A/G/C anchored primer to 1 ug/ul

Mix A, G and C-A/G/C anchored primers at ratio 1 ul: 1 ul, final concentration for A anchored and G anchored =0.5 ug/ul, final concentration for CA/CO/CC =0.5 ug/ul (0.167 ug /ul for each).

These concentrations will provide a theoretically equal probability for each primer to match the expressed sequences, assuming that each nucleotide A, G, C, and T in the last position before poly A sequence is equally distributed in the total expressed sequence pool.

d. Synthesis of first strand cDNA (Gibco cDNA synthesis kit number 18267-021)

items lx 2x 3x 4x 5x 5x first strand buffer 10 mM dNTP mix.
2.55 7.5101 2.5 RT oligo246810 RNAs in 12345 DTT (100mM) 51 01 52 0
25mM L V
RT2 .557 .5101 2.5 mRNA(1ug/ul) 510152025 H_2O *23466992105

Depending on the input of mRNA, the final volume can be adjusted with H_2O. The mix is incubated at 37°C for 30 minutes; 55°C 2 minutes; and again at 37°C. 2 ul Reverse Transcriptase is added and incubated for 30 minutes. These steps are repeated once or twice more. This strategy significantly increases the cDNA yield by repeating dissociation/association of the RT primers which induces correct annealing to the right position and increases the initiation of cDNA synthesis. The original quantitative relationship between each template is not altered within the total population, because the probability to be reverse transcribed is proportional to the original concentration. This is followed by removal of 5-10 ul from the reaction and the addition of 1 ul DNase-free RNase (Promega) and incubating at 37°C for 30 minutes. The product is assayed by gel electrophoresis and the remainder cDNA is aliquoted into tubes at a concentration of 50 ul/tube.

e. Synthesis of second strand cDNA

items lx 2x 3x 4x 5x DEPC H_2O_2 9058 087 011 6 014 5 O dNT P mix(10mM)
7 .5152 2 0.5303 7.51 x 2nd strand buffer 4080120160200 Eco1i DNA polymerase I
10 20 30 40 50 Ecoli DNA ligase I 252.5 3.755 6.25

Add 350 ul to each 1st strand cDNA tube and mix well. Incubate at 16°C for 2 hours and assay on a gel together with 1st strand cDNA. Treat the cDNA with 400 ul phenol/chloroform(25/24/1) and recover the upper phase. Aliquot cDNA in new tubes at a concentration of 300 ul/tube. Add 3 ul glycogen, 150 ul 7.5 M NH_4OAC, 600 ul cold ethanol. Mix well by vortexing and centrifuge for 15 min at maximal speed at room temperature. Combine all pellets into one tube and wash with 500 ul 70% ethanol. Dry the pellet and dissolve the pellet in 20-50 ul TE. Make double dilutions from 1:0 to 1:10 with 2 ul of cDNA and quantify the concentration by dot quantification with known standard dilutions of DNA. Adjust the concentration to 200 ng /ul. The efficiency of cDNA synthesis = ds cDNA/input mRNAxl00%. The typical value is about 1:1.7 (mRNA: ds cDNA)

3. NIaIII digestion of double strand cDNA

double strand cDNA	75 ul (-20 ug)
buffer 4 (NEB)	10 ul
10 x BSA (NEB)	10ul,
NIaIII	5 ul,

Incubate at 37°C for 1-2 hours and assay 1 ul of digest on a gel. The cDNA should be centered at about 300 - 500 bps

4. Recovery of 3' cDNA pool

Aliquot 200 ul Dynal M280 beads into a tube for each 4 ug cDNA and wash beads with 500 ul binding/washing buffer. Resuspend beads in 200 ul B/W and add 20 ul cDNA, 180 ul H_20. Rotate or shake at room temp for 20-60 minutes. Isolate the beads in microcentrifuge and save the supernatant. Wash the beads with 3x 500 ul 1 x B/W, 1 x 1000 ul TE and resuspend them in 50 ul TE and combine beads into one tube. Add equal volume of phenol, vortex 5 minutes at maximal speed and incubate at 65°C for 30 minutes. Repeat the vortex every 5 minutes and the vortex at full speed for 10 minutes . Centrifuge to recover upper phase. Perform a phenol-chloroform extraction and add 1/2 v. 7.5 M NH40AC, 2 v. 100% ethanol, and 3 ul glycogen and centrifuge for 15 min at room temperature. Combine all pellets into one tube and wash with 70% ethanol and dissolve the pellets in 20 ul TE.

This is followed by a Not I digestion using the following:

NEB buffer 3 +BSA	3
DNA	20
Not I (10 u/ul)	3

Incubating the above mixture at 37°C overnight, extracting with Phenol-chloroform and performing an ethanol precipitation. The final cDNA was diluted in 22 ul H_20 and 1 ul cDNA was assayed on an agarose gel.

5. Cloning the recovered 3' cDNA

The Not I/SphI digested pGEM5Zf(+) vector was prepared using the following:

pGEM5zf(+)	10 ug/10 ul
NEB buffer 3 + BSA	4
Not I (10 u/ul)	2
SphI (10 u/ul)	2
H_20	22

The vector was digest at 37°C for more than 2 hours and 1 ul CIP was added at 37°C and incubated for 1 hour. GenClean beads purification was performed following the manufacturer's instruction. Purified DNA was resuspend to 200 ng/ul. Inserts of the linearized vector were cloned by mixing the following:

linearized vector	1
insert	5
buffer	1
ligase	1

and incubating at 16°C over night. This is followed by transformation wherein 5 ul of ligation mixture is added to 50 ul JM109 and incubated on ice for 30 min. This is followed by a heat shock at 42°C for 2 min and cooling the cells on ice for 2 min. 800 ul SOC is added and the culture is incubated in a shaker at 250 rpm at 370C for 30 min. This is followed by plating 100 ul of transformants on plates containing IPTG/Xgal/Amp and incubating the plated overnight.

Clones are screened by picking clones directly in PCR mixture (T7/sp6primer) and performing a PCR. The PCR products are verified on an agarose gel and purified using a S-300 column. The PCR product are sequenced by a PE BigDye kit with SP6 primer and the sequence is used to determine the presence of dT16.

The positive clones are subject to a large scale preparation of plasmids starting with a 400 ml LB with 50 ug/ml amp culture that is incubated in a shaker at 250 rpm at 37°C overnight. The plasmids are prepared by Qiagen Maxi plasmid preparation kit and the recovered plasmid concentration is adjusted to 1 ug/ul.

6. Preparation of drivers for normalization reaction

Inserts are released suing the following reaction mixture:

plasmid	30
Not I	3
ApaII	3
H20	18

NEB buffer 3	6

Incubate the digestion overnight. Run the digestion in 1% agarose gel to separate inserts from vector. Cut the gel containing the inserts and purify the DNA by GenClean kit. Resuspend purified inserts in TE

7. Conversion of double strand to single strand cycle DNA

Gene II digestion is performed using the following reaction mixture:

plasmid	100 ug/50
Gene II buffer	0 8
Gene II	8
H_20	24

Incubate at 30°C for 1 hour and then at 65°C 10 min then again at 37°C.

Exon III digestion is performed by adding 8 ul Exon III to the sample, incubating at 37°C for 60 min and treating with 100 ul of phenol-chloroform twice. This is followed by precipitation of the DNA using the following mixture:

DNA	80
7.5 M NH_4OAC	40
Ethanol	200

Incubating on dry ice for 15 min, centrifuging the mixture and washing the pellet with 70% ethanol, drying the pellet and resuspending the pellet in 40 ul H_20.

PvuII digestion is performed by add to the DNA 5 ul NEB buffer number 2, 5 ul pVUII and incubating at 37°C for 2 hrs

This is followed by the removal of double strand (ds) DNA by aliquoting 500 ul HAP (pH 6.8) in a tube and incubating at 65°C. The PvuII digested DNA is resuspended in HAP by vortexing and allowing the binding reaction to proceed for 1 min. This is followed by centrifuging the mix for 10 sec and recovering 400 ul supernatant. The single strand (ss) DNA is then desalted by adding the DNA to Sephadex G50 minicolumn (20ul/column), centrifuging and collecting the elutes together (about 17 ul/each, total about 370 ul). The precipitation of the purified ss DNA is performed using the reaction mix described below:

ss DNA elutes	350
7.5 M NH_4OAC	150
Ethanol	875
Glycogen	2

Incubating this mixture on dry ice for 20 min, centrifuging, washing and resuspending the ssDNA in 20 ul H20. 1 ul of ssDNA is run on a gel and quantified at OD260 and the final concentration is adjusted to 100 ng/ul.

8. Normalization is performed using the reaction mix below:

ssDNA 100 ng/ul	1
insert 100 ng/ul	4
4x hyb buffer	2
H20	1

This is incubated at 98°C for three minutes and then at 68°C overnight.

HAP absorption to remove the hybrids is performed by diluting HAP by adding 50 ul PB (pH6.8) to 50 ul HAP, incubating HAP at 60°C and resuspending the normalization mixture to HAP and centrifuging to recover the supernatant. The supernatant is then passed through a Sephadex G50 minicolumn twice to desalt. All the elutes are pooled and precipitated with 7.5 M NH_40AC, as before. The purified DNA is resuspend in 22 ul H_20 and the DNA is assayed on a gel as before.

9. Conversion of single strand to double strand

Sp6 primer is chosen because of the orientation of fl origin in this vector. The reaction mixture is as follows:

ssDNA	11
5x Sequenase buffer	4
Sp6 primer (1 ug/ul)	1

and is incubated at 65°C for 5 min. This is followed by incubating at 30°C for 5 min, and adding the following:

Sequenase	1
DTT 100mM	1
DNTP 10 mM	2

The mixture is then incubated at 30°C for 2 min followed by at 37°C for 30 min. A phenol chloroform extraction is performed followed by precipitation with glycogen dissolved as described before.

10. Preparation of plasmid

Plasmids are prepared by transformation as described before followed by mini preparation of clones with Qiagen REAL system following the manufacturer's protocol.

11. Sequence collection

Each clone is sequenced with a T7 primer (5' reading through), or Sp6 primer (3' reading through). T7 generates better sequence, but cannot be used if a SAGE Tag is in the sequence as 22 bps downstream of the primer is too near, Sp6 can detect the anchored dT16 sequence and generate SAGE Tags for many of the clones, but the sequencing quality is affected with more N after the dT16.

The sequencing sample can be cleaned and precipitated by 3M NaOAc precipitation. The sequences are then collected in an ABI377 auto sequencer, the sequence quality is checked and proximal linker sequences and distal linker sequence after CATG are removed.

12. Gene identification

The genes are identified by performing a BLAST on each sequence to the NCBI databases. First, the procedure involves aligning the sequence with NR, if a match is detected the gene is a known gene; if no match is detected the sequence is aligned with human EST. If a match is detected the gene is a known human EST sequence and if no match is found in both the NR and EST it is concluded that a novel gene is found.

Alternatively a SAGE Tag databases may be used. However, the currently available SAGE tag are limited to certain tissues (colon, pancreas, brain etc). If the targeted samples are not these tissues, the SAGE databases will not be very useful.

The SAGE tag sequences can also be used as probes to match the standard databases. This can provide an independent confirmation for a particular sequence.

References

The following references, to the extent that they provide exemplary procedural or other details supplementary to those set forth herein, are specifically incorporated herein by reference.

Abbotts *et al.*, *J. Biol. Chem.* 266:3937-3943, 1991.

Adler and Modrich, *J. Biol. Chem.*, 254:11605-11614, 1979.

Alberts *et al.*, *In: Molecular Biology of the Cell*, Robertson (Ed.), Garland, New York, p 369, 1994.

An *et al.*, *J Clin Microbiol*;33(4):860-867 1995.

Bakhanashvili and Hizi, *FEBS Lett.* 319:201-205, 1993.

Beaucage, and Lyer, *Tetrahedron*, 48:2223-2311, 1992

Bebenek and Kunkel, *Nucl. Acids Res.*, 17:5408, 1989.

Belkin and Jannasch, *Arch. Microbiol.*, 141:181-186, 1985.

Bertling *et al.*, *PCR Methods Appl.*, 3:95-99, 1993.

Bhattacharyya *et al.*, *J. Biol. Chem.*, 270:1705-1710, 1995.

Boguski, *Trends Biochem. Sci.*, 21:295-296, 1995.

Bonaldo *et al.*, *Genome Res.* 6:791-806, 1996.

Butler and Chamberlin, *J. Biol. Chem.*, 257:5772-5778, 1982.

Chen *et al.*, *Mol Med.* 1(2): 153-160, 1995.

Chomczynski and Sacchi, *Anal Biochem.* 162(1): 156-159, 1987

Colgan and Manley, *Genes. Dev.* 11:2755-2766, 1997.

D'Alessio and Gerard, *Nucl. Acids Res.*, 16:1999-2014, 1988.

Dale *et al.*, *Plasmid*, 13:31-40, 1985.

Davanloo *et al.*, *Proc. Nat'l Acad. Sci. USA*, 81:2035-2039, 1984.

Davey *et al.*, EP No. 329 822

Derbyshire *et al.*, *Science*, 240:199-201, 1988.

DeRisi *et al.*, *Nat. Genet.*, 14:57-460, 1996.

Diatchenko *et al.*, *Proc. Nat'l Acad. Sci. USA*, 93:6025-6030, 1996.

Donahue *et al.*, *J. Biol. Chem.* 269: 8604-8609, 1994.

Duguid and Dinauer, *Nucl. Acids Res.*, 18:2789-2792, 1990.

Dwomiczak and Milault, *Nucl. Acids Res.*, 15:5181-5197, 1987.

Eckert and Kunkel, *PCR Methods and Applications*, 1:17-24, 1991.

Engler *et al.*, *J. Biol. Chem.*, 258:11165-11173, 1983.

Freifelder, *Physical Biochemistry Applications to Biochemistry and Molecular Biology*, 2nd ed. Wm. Freeman and Co., New York, NY, 1982.

Frohman, In: PCR PROTOCOLS: A GUIDE TO METHODS AND APPLICATIONS, Academic Press, N.Y., 1990;

Gerhold and Caskey, *BioEssays*, 18:973-981, 1996.

Gillam *et al.*, *J. Biol. Chem.* 253, 2532, 1978.

Gillam *et al.*, *Nucleic Acids Res.* 6, 2973, 1979.

Gingeras *et al.*, PCT Application WO 88/10315

Green *et al.*, *Cell*, 32:681-694, 1983.

Grippo and Richardson, *J. Biol. Chem.*, 246:6867-6873, 1971.

Gubler and Hoffmann, *Gene*, 25:263-269, 1983.

Gubler, *Methods Enzymol.*, 152:330-335, 1987.

Hori *et al.*, *J. Biol. Chem.*, 254:11598-11604, 1979.

Houts *et al.*, *J. Virol.*, 29:517-522, 1979.

Hugh and Griffin, *PCR Technology*, 228-229, 1994.

Iiyy *et al.*, *Biotechnique* 11:464, 1991.

Innis *et al.*, *PCR™ Protocols*, Academic Press, Inc., San Diego CA, 1990.

Itakura and Riggs, *Science* 209:1401-1405, 1980.

Itakura *et al.*, *J. Biol. Chem.* 250, 4592 1975

Ivanova and Belyavsky, *Nucl. Acid Res.*, 23:2954-2958, 1995.

Jannasch *et al.*, *Applied Environ. Microbiol.*, 58:3472-3481, 1992.

Karl, G., *Cell and Molecular Biology*. Page 467. John Wiley and Sons, USA, 1996.

Kato, *Nucl. Acid Res.*, 23:3685-3690, 1995.

Khorana, *Science* 203, 614 1979

Kong *et al.*, *J. Biol. Chem.*, 268:1965-1975, 1993.

Krieg and Melton, *Nucl. Acids Res.*, 12:7057-7070, 1984.

Kunkel *et al.*, *Methods Enzymol.*, 154:367-382, 1987.

Kwoh *et al.*, *Proc. Nat. Acad. Sci. USA*, 86: 1173, 1989.

Lehman, *In: The Enzymes*, Boyer (Ed.), Vol. 14A, pp 16-38, Academic Press, San Diego, CA, 1981.

Liang and Pardee, *Science*, 257:967-970, 1992.

Liang *et al. Nucleic Acids Res.* 22:5763-5764, 1994.

Lockhart *et al.*, *Nature Biotech.*, 14:1675-1680, 1996.

Maniatis *et al.*, *Cell*, 8:163, 1976.

Mattila *et al.*, *NAR*, 19:4967-4973, 1991.

McClary *et al.*, *J. DNA Sequencing Mapping*, 1(3): 173-180, 1991.

Mead *et al.*, *BioTechniques*, 11(1): 76-87, 1991.

Meinkoth and Wahl, *Methods Enzymol.*, 152:91-94, 1987.

Melton, *Proc. Nat'l Acad. Sci. USA*, 82:144-148, 1985.

Miller *et al.*, PCT Application WO 89/06700

Modrich and Richardson, *J. Biol. Chem.*, 250:5515-5522, 1975.

Mok *et al.*, *Gynecol Oncol.* 52(2): 247-252, 1994

Murray and Kelley, *Molec. Gen. Genet.*, 175:77-87, 1979.

Nordstrom *et al.*, *J. Biol. Chem.*, 256:3112-3117, 1981.

Noren, *Nucl. Acids Res.*, 18:83-88, 1990.

Ohara *et al.*, *Proc. Nat'l Acad. Sci. USA*, 86: 5673-5677, 1989.

Perler *et al.*, *Adv. Protein Chem.* 48:377-435, 1996

Perler *et al.*, *Proc. Nat'l Acad. Sci. USA*, 89:5577, 1992.

Ponte *et al.*, *Nucl. Acids Res.*, 12:1687-1696, 1984.

Prashar and Weissman, *Proc. Natl. Acad. Sci. USA*, 93:659-663, 1996.

Promega: 1993. Protocols and Applications Guide (2nd edition), p58-61, Promega, Madison, USA.

Sager *et al.*, FASEB J. 7(10): 964-970, 1993.

Sambrook *et al.*, *In: Molecular Cloning: A Laboratory Manual*, second edition, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1989.

Sanger *et al.*, *Proc. Nat'l Acad. Sci. USA*, 74:5463-5467, 1977.

Schenborn and Meirendorf, *Nucl. Acids Res.*, 13:6223-6236, 1985.

Siebert *et al.*, *Nuc. Acids. Res.* 23:1087-1088, 1995.

Strausberg *et al.*, *Nat. Genet.*, 15:415-416, 1997.

Studier *et al.*, *Methods Enzymol.*, 185:60-89, 1990.

Sun *et al.*, *Cancer Res.* 54:1139-1144, 1994.

Tabor and Struhl, *In: Current Protocols in Molecular Biology*, Ausubel *et al.* (Eds.), John Wiley and Sons, NY, pp 3.5.10-3.5.12, 1989.

Tanese and Goff, *Proc. Nat'l Acad. Sci. USA*, 85:1977, 1988.

U. S. Patent 4,704,362

U. S. Patent 5,221,619

U. S. Patent 5,583,013

U.S. Patent 4,659,774

U.S. Patent 4,683,195,

U.S. Patent 4,683,202

U.S. Patent 4,800,159,

U.S. Patent 4,816,571

U.S. Patent 4,883,750

U.S. Patent 4,959,463

U.S. Patent 5,141,813

U.S. Patent 5,262,311

U.S. Patent 5,264,566

U.S. Patent 5,428,148

U.S. Patent 5,554,744

U.S. Patent 5,574,146

U.S. Patent 5,602,244

U.S. Patent 5,665,547

Velculescu *et al.*, *Science*, 270:484-487, 1995.

Walker *et al.*, *Proc. Nat'l Acad. Sci. USA*, 89:392-396 1992.

Watson *et al.*, *Cancer Res.* 54(17): 4598-4602, 1994

Welsh *et al. Nucleic Acids Res.* 20(19): 4965-4970, 1992

Wu *et al.*, *Genomics*, 4:560, 1989.

Yu and Goodman, *Biol. Chem.* 267:10888-10896, 1992.

Zhang *et al.*, *Science*, 276:1268-1272, 1997.

Zinn *et al.*, *Cell*, 34:865-879, 1983.

Claims:

1. A method for amplifying a first set of target polynucleotides containing poly-A sequences comprising:

(a) providing a set of five primers, wherein each of said primers is comprised a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC;

(b) annealing said primers to said first set of target polynucleotides;

(c) contacting said primer-annealed first set of target polynucleotides with a polymerase and dNTPs; and

(d) subjecting the components of step (c) to conditions permitting polymerization,

whereby a first set of polymerization products is generated.

2. The method of claim 1, wherein said polymerase is a reverse transcriptase.

3. The method of claim 2, wherein said reverse transcriptase is MMLV reverse transcriptase.

4. The method of claim 1, wherein said polymerase is a DNA polymerase.

5. The method of claim 4, wherein said DNA polymerase is Taq.

6. The method of claim 1, wherein said primers further comprise a sequence encoding a promoter 5′ to the poly-dT sequence.

7. The method of claim 6, wherein said promoter is an SP6 promoter, an M13 promoter, a T3 promoter or a T7 promoter.

8. The method of claim 1, further comprising subjecting said first set of polymerization products to PCR.

9. The method of claim 1, further comprising separating said first set of polymerization products.

10. The method of claim 9, wherein separating comprises gel electrophoresis.

11. The method of claim 10, wherein gel electrophoresis comprises denaturing gel electrophoresis.

12. The method of claim 1, wherein said poly-dT sequence is about 10 to about 35 bases.

13. The method of claim 11, wherein said poly-dT sequence is 11 bases.

14. The method of claim 1, wherein said primers contain a label.

15. The method of claim 14, wherein said label is a fluorometric label, colorimetric label, enzymatic label, magnetic label, biotin label or radioactive label.

16. The method of claim 1, wherein said target polynucleotide is an RNA.

17. The method of claim 1, wherein said target polynucleotide is a DNA.

18. The method of claim 1, wherein said first set of polymerization products are compared with a second set of polymerization products generated from a second set of target polynucleotides.

19. The method of claim 18, wherein said comparison is differential display.

20. A method for generating DNA library from poly-A RNAs comprising:

(a) providing a set of five primers, wherein each of said primers is comprised a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC;

(b) annealing said primers to said RNAs;

(c) contacting said primer annealed set of target polynucleotides with a polymerase dNTPs;

(d) subjecting the components of step (c) to conditions permitting polymerization; and

(e) cloning polymerization products of step (d) into a suitable vector;

whereby a DNA library is generated.

21. The method of claim 20, wherein said primers further comprise a sequence encoding a promoter 5′ to the poly-dT sequence.

22. The method of claim 21, wherein said promoter is a SP6 promoter, a T3 promoter or a T7 promoter.

23. The method of claim 20, wherein said vector is an expression vector.

24. The method of claim 20, wherein said polymerase is a reverse transcriptase.

25. The method of claim 20, further comprising subjecting said polymerization products to PCR.

26. A method for performing differential display comprising:

(a) providing a set of five primers, wherein each of said primers is comprised a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC;

(b) annealing said primers to a first set of target polynucleotides containing poly-A sequences;

(c) contacting said primer-annealed first set of target polynucleotides with a polymerase and dNTPs;

(d) subjecting the components of step (c) to conditions permitting polymerization to create a first set of polymerization products; and

(e) comparing said first set of polymerization products with a second set of polymerization products produced according to steps (a)-(d) using a second set of target polynucleotides containing poly-A sequences.

27. The method of claim 26, further comprising subjecting said first set of polymerization products to PCR.

28. A kit comprising five poly-dT primers, wherein each of said primers comprises, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC.

29. The kit of claim 28, further comprising a polymerase.

30. The kit of claim 29, wherein said polymerase is a reverse transcriptase.

31. The kit of claim 29, wherein said polymerase is a DNA polymerase.

32. The kit of claim 28, wherein at least one of said primers comprises a label.

33. The kit of claim 32, wherein each of said primers comprises a distinct label.

34. The kit of claim 32, wherein said label is a fluorometric label, colorimetric label, enzymatic label, biotin label, magnetic label or radioactive label.

35. The kit of claim 28, further comprising standard polynucleotides suitable for amplification by each of said primers.

36. The kit of claim 28, wherein said poly-dT sequence is about 10 to about 35 bases.

37. The kit of claim 36, wherein said poly-dT sequence is 11 bases.

38. The kit of claim 37, further comprising arbitrary primers.

39. A method for identifying an expressed gene fragment comprising:

(a) providing a polyA-minus cDNA population labeled at its 3'-end;
(b) digesting said cDNA population with a restriction enzyme;
(c) isolating the 3' fragments of said population;
(d) performing 3' cDNA subtraction on said fragments;
(e) performing suppression PCR on said subtracted fragments; and
(f) identifying a gene fragment from said amplified fragments.

40. The method of claim 39, further comprising reverse transcribing an mRNA population into said cDNA population.

41. The method of claim 40, wherein said label is biotin.

42. The method of claim 40, wherein the primers used for reverse transcription consist of a poly-dT sequence and, at the 3' end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of A, G, CA, CG and CC.

43. The method of claim 42, wherein the primers further comprise the sequence TTTGCATGCTCGAG 5' to said poly-dT sequence.

44. The method of claim 42, wherein said poly-dT sequence is about 10 to about 35 bases.

45. The method of claim 44, wherein said poly-dT sequence is 16 bases.

46. The method of claim 39, wherein said restriction enzyme is *Nla*III.

47. The method of claim 39, further comprising verifying the subtraction efficiency.

48. The method of claim 49, wherein verification is via multiplex quantitative PCR.

49. The method of claim 48, wherein targets for said PCR are one or more of the β-actin gene, the *HSC*70 gene and the *HSP*75 gene.

50. The method of claim 39, further comprising cloning said isolated gene fragment.

51. The method of claim 50, further comprising sequencing of said cloned gene fragment.

52. The method of claim 51, wherein said sequencing is one-pass sequencing.

53. The method of claim 51, wherein said sequencing is SAGE sequencing.

54. The method of claim 52, further comprising comparing the resulting sequence with one or more sequencing-containing databases.

55. The method of claim 52, further comprising identifying a plasmid containing the matched sequence from the I.M.A.G.E. consortium.

56. The method of claim 50, further comprising probing a cDNA library with said cloned gene fragment.

57. The method of claim 56, further comprising isolating a complete cDNA corresponding to said cloned gene fragment.

58. The method of claim 57, further comprising cloning said complete cDNA.

59. The method of claim 58, further comprising sequencing said cloned complete cDNA.

60. A method for identifying an expressed gene fragment comprising:

(a) converting mRNA molecules into a polydA/dT-minus cDNA population;

(b) digesting said cDNA population with a restriction enzyme;

(c) isolating the 3' DNA fragments of said population thereby generating a 3' polydA/dT-minus cDNA library;

(d) generating from said cDNA library

(i) a single-stranded cDNA library, and

(ii) double-stranded inserts,

(e) performing a subtraction on said single-stranded library using said double-stranded inserts;

(f) eliminating double-stranded hybrids, thereby isolating a circular single-stranded cDNA sublibrary; and

(g) sequencing the cDNA clones from step (f).

61. The method of claim 60, further comprising prior to step (a) the step of obtaining mRNA molecules.

62. The method of claim 60, wherein said converting mRNA molecules into said cDNA population comprises the use of anchored dT primers, a polymerase and dNTPs.

63. The method of claim 62, wherein said anchored polydT primers are each comprised of a poly-dT sequence and, at the 3′ end of the poly-dT sequence, a nucleic acid singlet or doublet selected from the group consisting of dA, dG, CA, CG and CC.

64. The method of claim 62, wherein said poly-dT sequence is about 10 to about 35 bases.

65. The method of claim 62, wherein said poly-dT sequence is 16 bases.

66. The method of claim 62, wherein said polymerase is a reverse transcriptase.

67. The method of claim 66, wherein said reverse transcriptase is MMLV reverse transcriptase.

68. The method of claim 66, wherein said reverse transcriptase is AMV reverse transcriptase.

69. The method of claim 60, wherein said restriction enzyme is NlaIII.

70. The method of claim 60, wherein said generation of polydA/dT- minus 3' cDNA library comprises cloning the isolated 3' cDNA fragments.

71. The method of claim 60, wherein said sequencing is one-pass sequencing.

72. The method of claim 60, wherein said sequencing is SAGE sequencing.

73. The method of claim 60, further comprising comparing the sequence obtained with one or more sequence-containing databases.

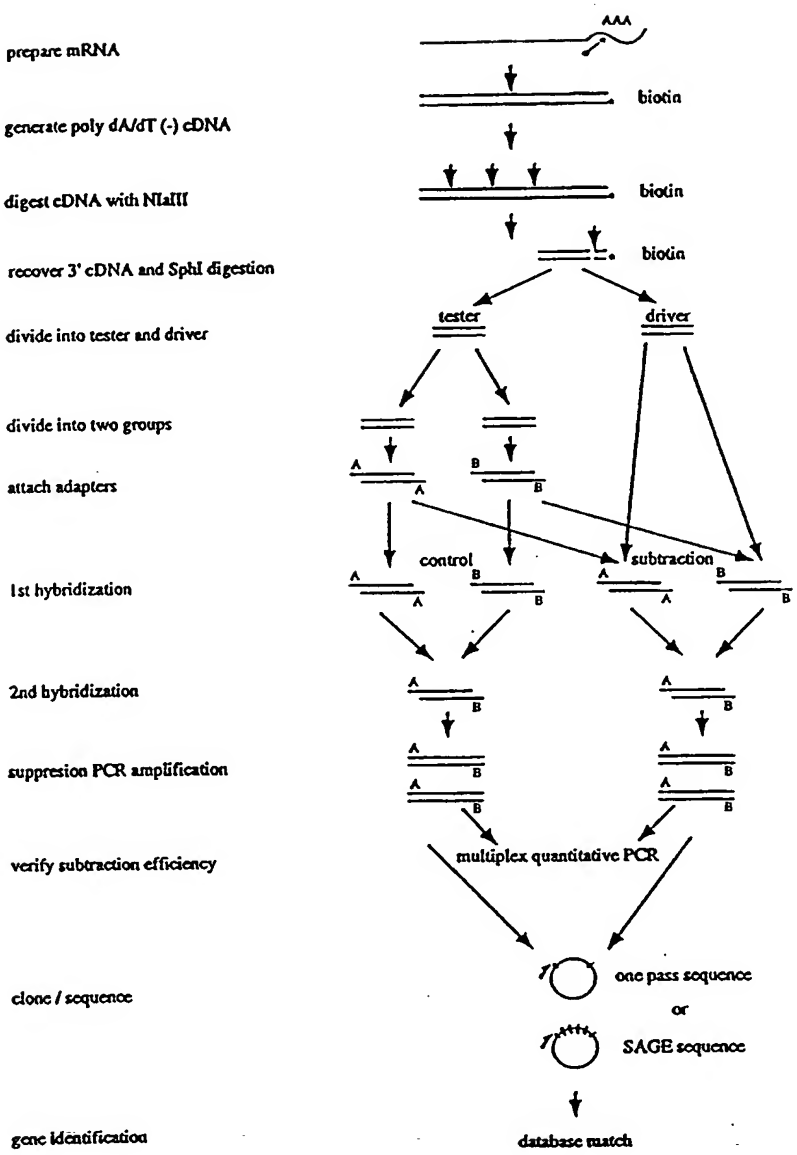

prepare mRNA
AAA
generate poly dA/dT (-) cDNA
biotin
digest cDNA with NlaIII
biotin
recover 3' cDNA and SphI digestion
biotin
divide into tester and driver
tester
driver
divide into two groups
attach adapters
A
B
A
B
1st hybridization
control
subtraction
2nd hybridization
suppresion PCR amplification
verify subtraction efficiency
multiplex quantitative PCR
clone / sequence
one pass sequence
or
SAGE sequence
gene identification
database match

FIG. 1

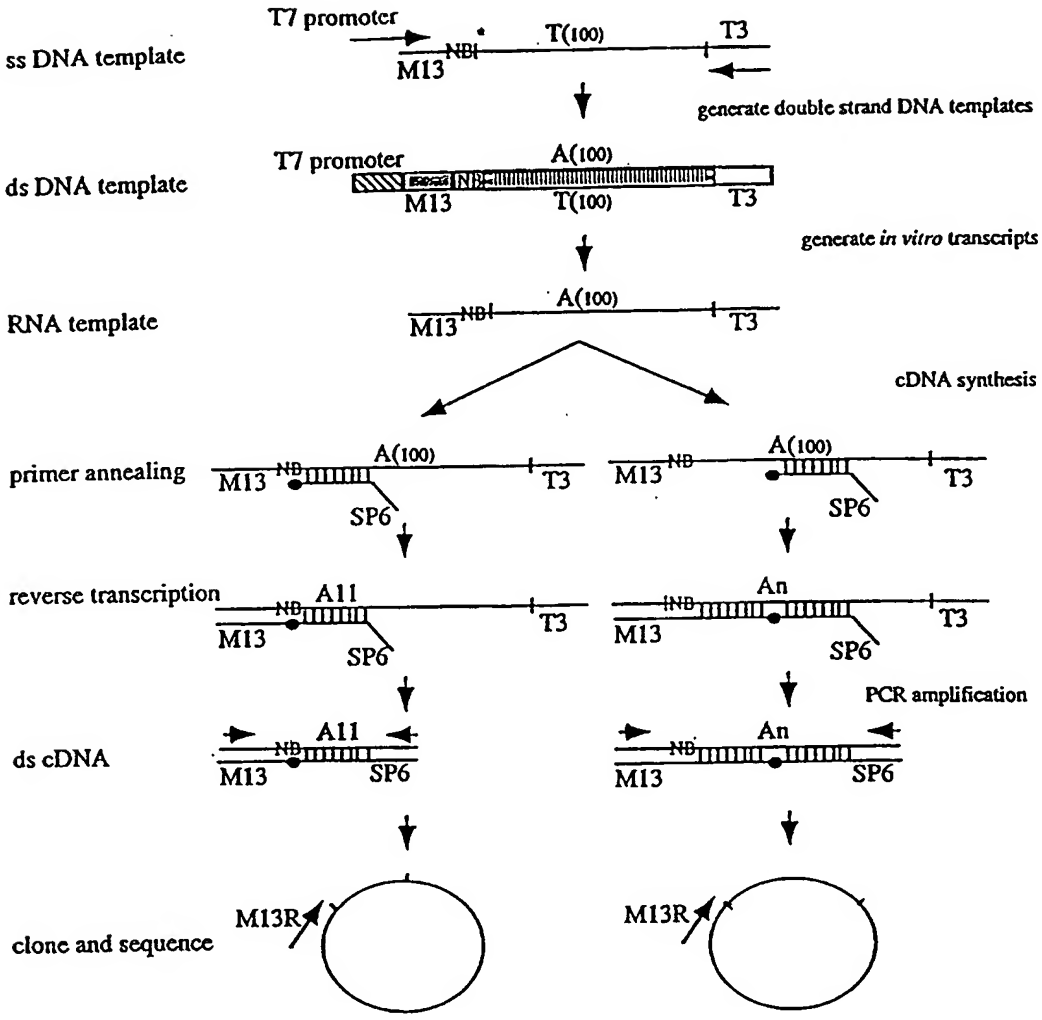
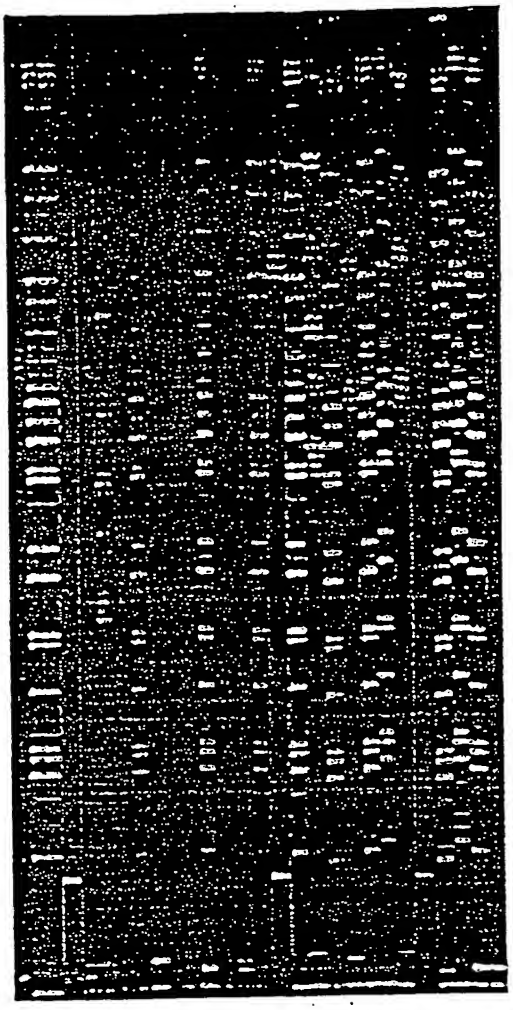

ss DNA template
T7 promoter
T(100)
T3
M13
NB
generate double strand DNA templates
ds DNA template
T7 promoter
A(100)
M13
T(100)
T3
generate *in vitro* transcripts
RNA template
A(100)
M13
NB
T3
cDNA synthesis
primer annealing
A(100)
M13
NB
T3
SP6
reverse transcription
A11
An
PCR amplification
ds cDNA
clone and sequence
M13R
* B=A, G, or C
N=A, G, C, T
anchored primers

FIG. 2



A11 or T11 indicate the correct products

FIG. 3

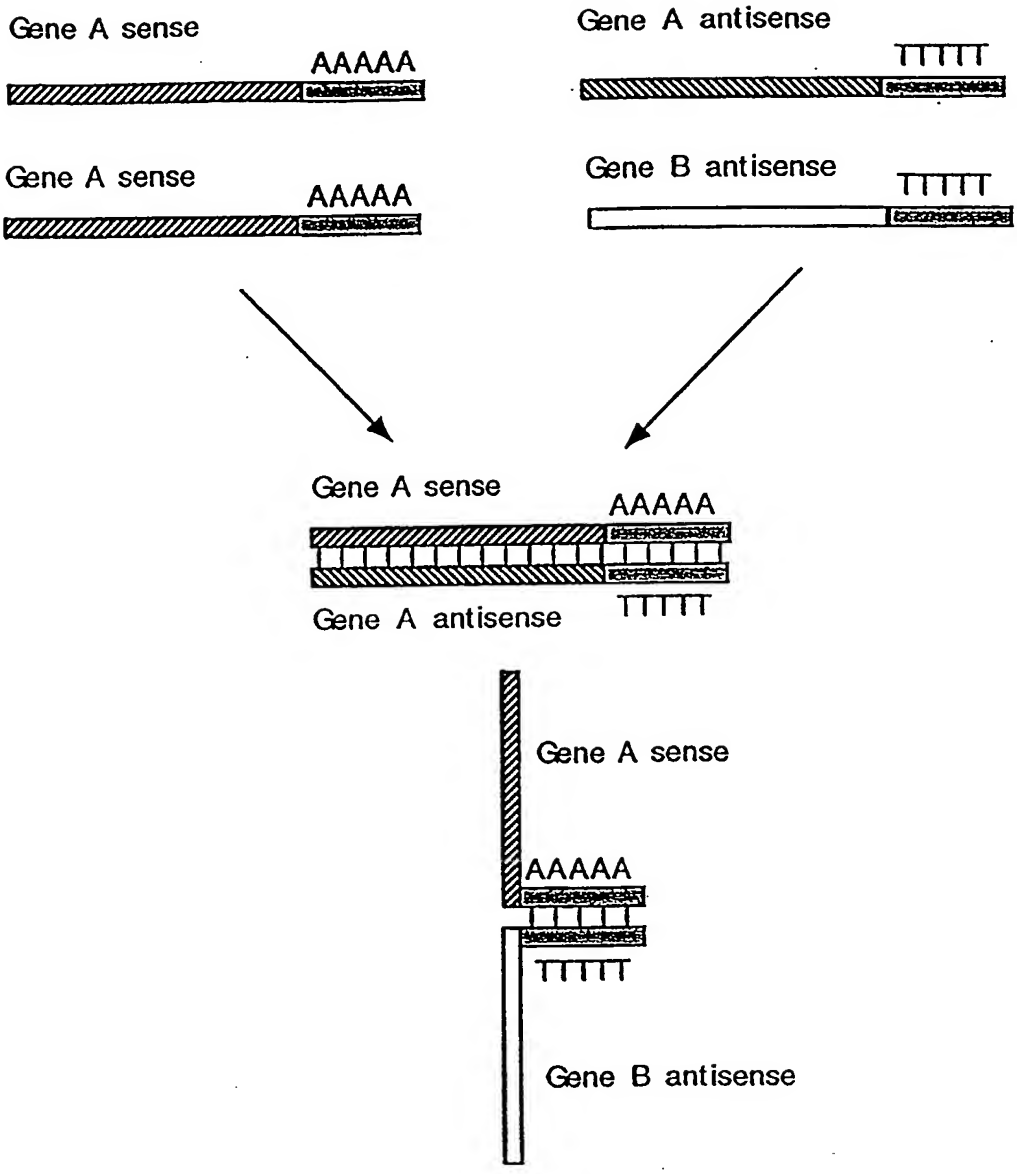

Gene A sense
AAAAA
Gene A antisense
TTTTT
Gene A sense
AAAAA
Gene B antisense
TTTTT
Gene A sense
AAAAA
Gene A antisense
TTTTT
Gene A sense
AAAAA
TTTTT
Gene B antisense

FIG. 4

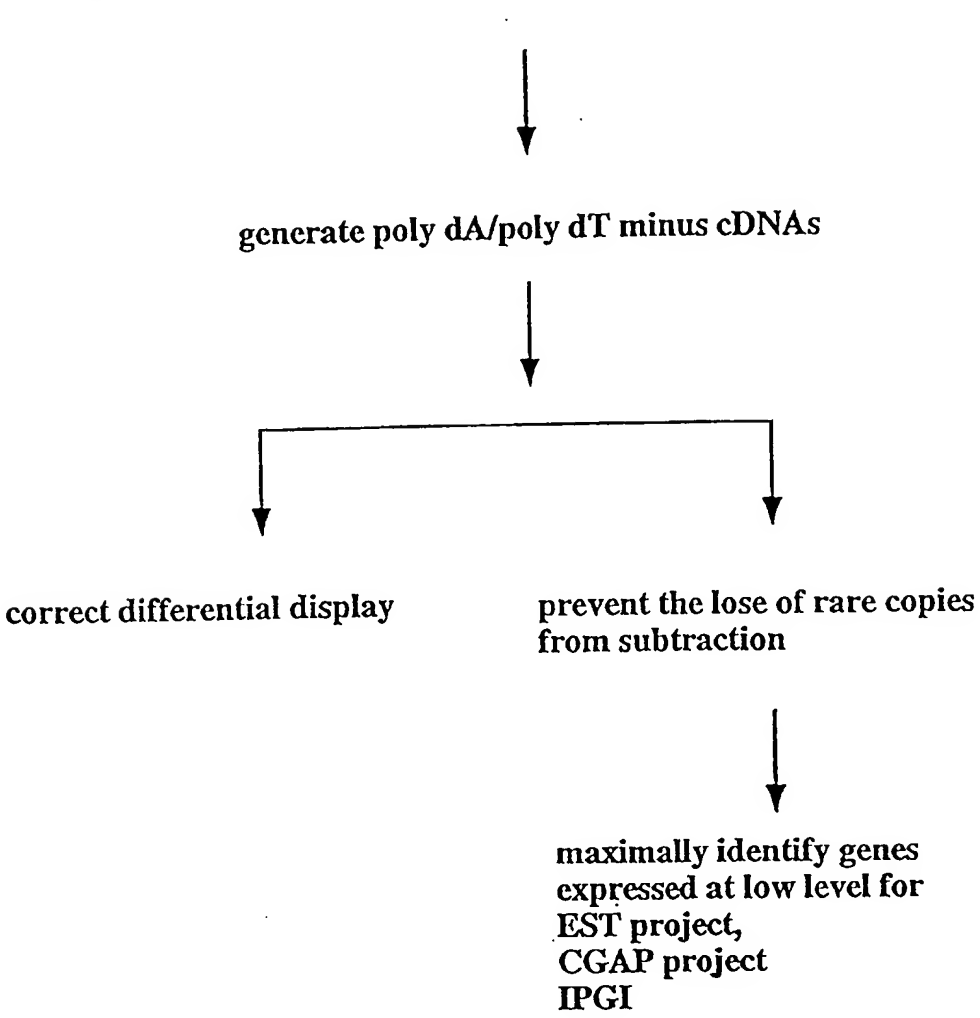

Use 5 different anchored oligo dT primers for reverse transcription
generate poly dA/poly dT minus cDNAs
correct differential display
prevent the lose of rare copies from subtraction
maximally identify genes expressed at low level for EST project, CGAP project IPGI

FIG. 5

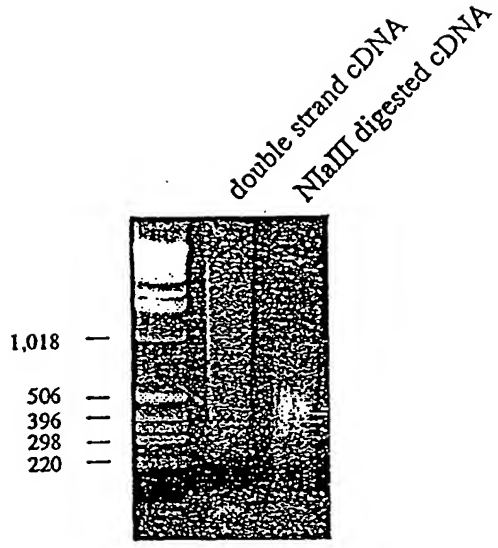

double strand cDNA
NlaIII digested cDNA
1,018
506
396
298
220

FIG. 6

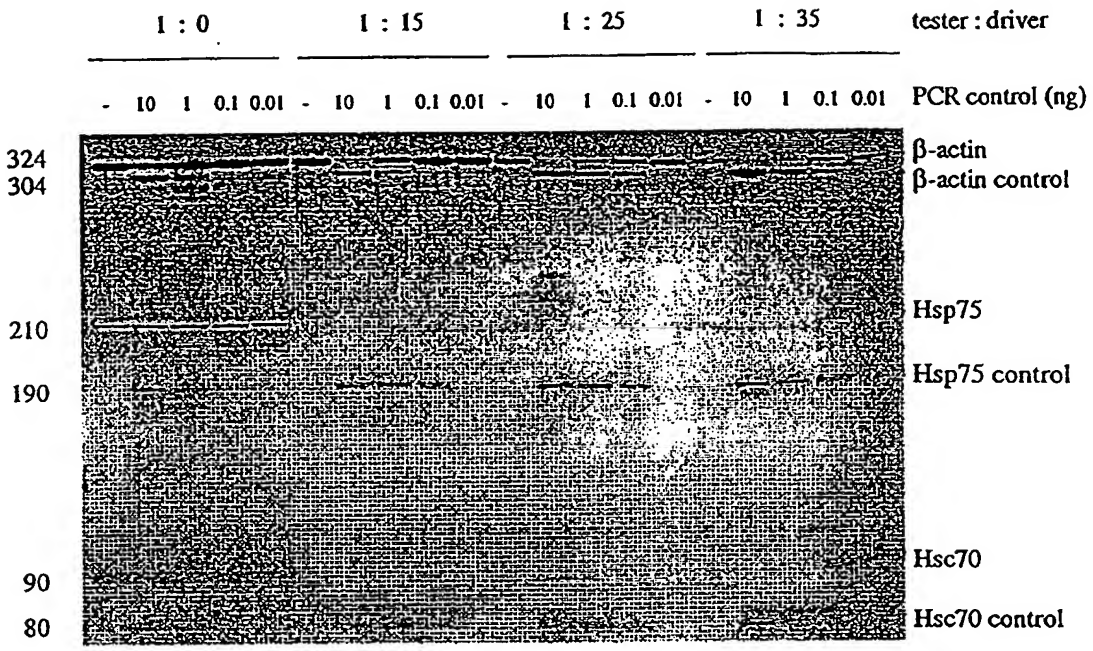

1 : 0
1 : 15
1 : 25
1 : 35
tester : driver
- 10 1 0.1 0.01 - 10 1 0.1 0.01 - 10 1 0.1 0.01 - 10 1 0.1 0.01
PCR control (ng)
324
304
210
190
90
80
β-actin
β-actin control
Hsp75
Hsp75 control
Hsc70
Hsc70 control

FIG. 7

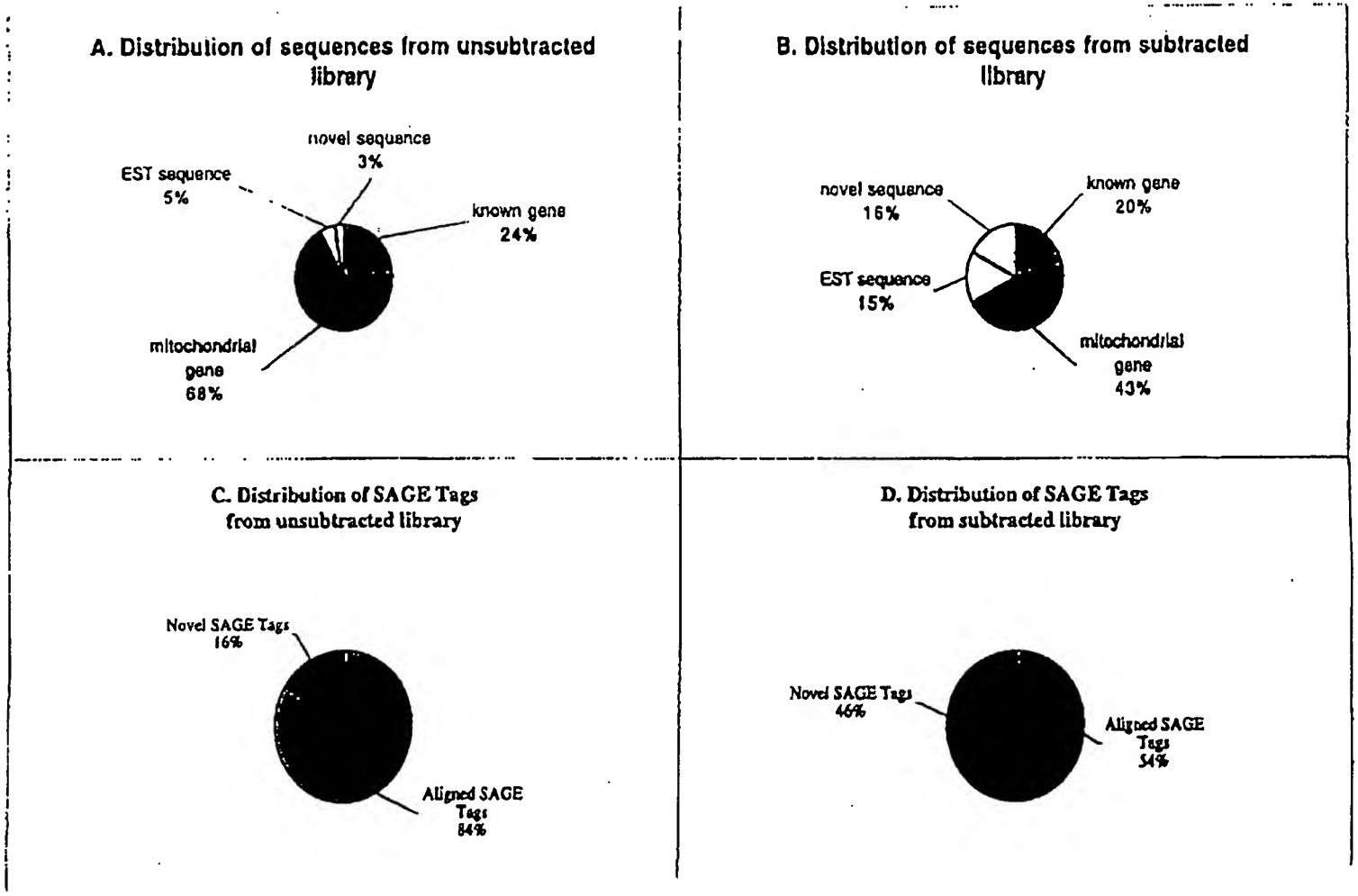
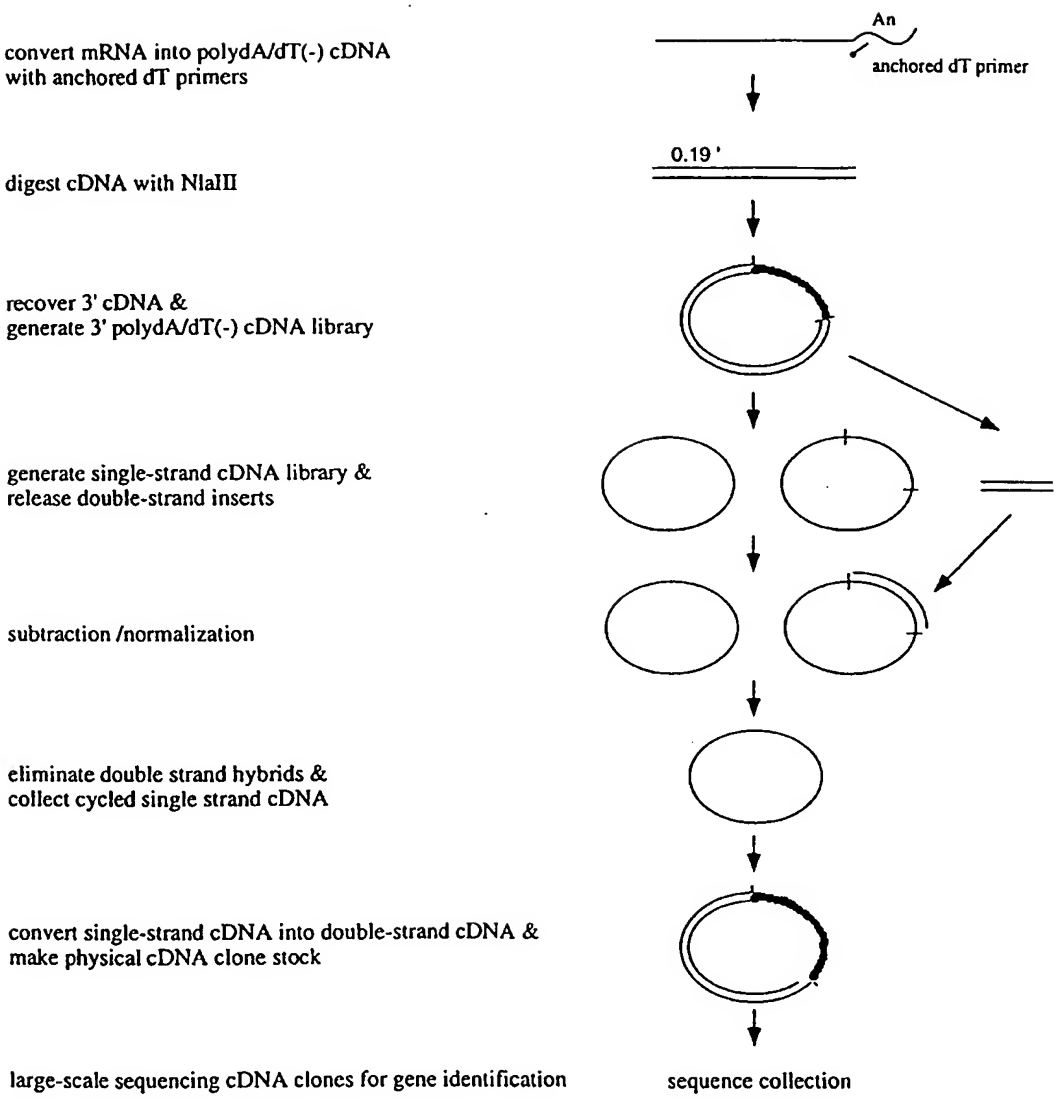

A. Distribution of sequences from unsubtracted library
novel sequence 3%
EST sequence 5%
known gene 24%
mitochondrial gene 68%
B. Distribution of sequences from subtracted library
novel sequence 16%
known gene 20%
EST sequence 15%
mitochondrial gene 49%
C. Distribution of SAGE Tags from unsubtracted library
Novel SAGE Tags 16%
Aligned SAGE Tags 84%
D. Distribution of SAGE Tags from subtracted library
Novel SAGE Tags 46%
Aligned SAGE Tags 54%

FIG. 8A – 8 D



Schematic of Screening polydA/dT(-) cDNAs for Gene Identification

FIG. 9

SEQUENCE LISTING

<110> Wang, San Ming
Fears, Scott
Rowley, Janet D.

<120> A NEW STRATEGY FOR GENOME-WIDE GENE ANALYSIS:
INTEGRATED PROCEDURES FOR GENE IDENTIFICATION

<130> ARCD:295 and ARCD:295P

<140> UNKNOWN
<141> 1999-09-29

<150> 60/102,381
<151> 1998-09-29

<170> 31

<180> PatentIn Ver. 2.0

<210> 1
<211> 142
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
primer

<220>
<221> modified_base
<222> (21)
<223> residue 21 = N = A, or G, or C, or T

<220>
<221> modified_base
<222> (22)
<223> residue 22 = B = A, or G, or C

<400> 1
gtaaaacgac ggccagtacg nbaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa 60
aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa aaaaaaaaaa 120
aactttagtg agggttaatt tc 142

<210> 2
<211> 42
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic
primer

<400> 2
cgtaatacga ctcactatag gggtaaaacg acggccagta cg 42

<210> 3
<211> 20
<212> DNA

<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 3
gaaattaacc ctcactaaag 20

<210> 4
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 4
ctaatacgac tcactatagg gc 22

<210> 5
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 5
cgatttaggt gacactatag 20

<210> 6
<211> 17
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 6
caggaaacag ctatgac 17

<210> 7
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 7
gtaaaacgac ggccagtacg 20

<210> 8
<211> 32
<212> DNA

<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 8
cgatttaggt gacactatag ttttttttt ta 32

<210> 9
<211> 32
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 9
cgatttaggt gacactatag ttttttttt tg 32

<210> 10
<211> 32
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 10
cgatttaggt gacactatag ttttttttt tc 32

<210> 11
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 11
cgatttaggt gacactatag ttttttttt tca 33

<210> 12
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 12
cgatttaggt gacactatag ttttttttt tcg 33

<210> 13
<211> 33
<212> DNA

<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 13
cgatttaggt gacactatag ttttttttt tcc 33

<210> 14
<211> 33
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 14
cgatttaggt gacactatag ttttttttt tct 33

<210> 15
<211> 31
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 15
tttgcatgct cgagtttttt ttttttttt a 31

<210> 16
<211> 31
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 16
tttgcatgct cgagtttttt ttttttttt g 31

<210> 17
<211> 31
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 17
tttgcatgct cgagtttttt ttttttttt c 31

<210> 18
<211> 43
<212> DNA

<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 18
atacgactca ctatagggct cgagcggccg catatgggac atg 43

<210> 19
<211> 10
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 19
tcccatatgc 10

<210> 20
<211> 43
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 20
atacgactca ctatagggca gctcgccggc gtatagggac atg 43

<210> 21
<211> 10
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 21
tccctatacg 10

<210> 22
<211> 40
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 22
tgttacagga agtcccttgc ttctctctaa ggagaatggc 40

<210> 23
<211> 20
<212> DNA

<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 23
tgttacagga agtcccttgc 20

<210> 24
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 24
taaggtgtgc acttttattc 20

<210> 25
<211> 35
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 25
ccaggaggaa tgcctggggt ggtggagctc ctcct 35

<210> 26
<211> 18
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 26
ccaggaggaa tgcctggg 18

<210> 27
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 27
ttaatcaacc tcttcaatgg 20

<210> 28
<211> 40
<212> DNA

<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 28
agataaaggc acaagacgtg tcttctggtg gattaagcaa 40

<210> 29
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 29
agataaaggc acaagacgtg 20

<210> 30
<211> 20
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 30
gcaggtaatt ggtccttgaa 20

<210> 31
<211> 22
<212> DNA
<213> Artificial Sequence

<220>
<223> Description of Artificial Sequence: Synthetic primer

<400> 31
ctaatacgac tcactatagg gc 22